

Unit 4

Surveys

Introduction

Units 1–3 have been largely concerned with stage 3 of the modelling diagram (shown in Figure 1), the analysis of the data.

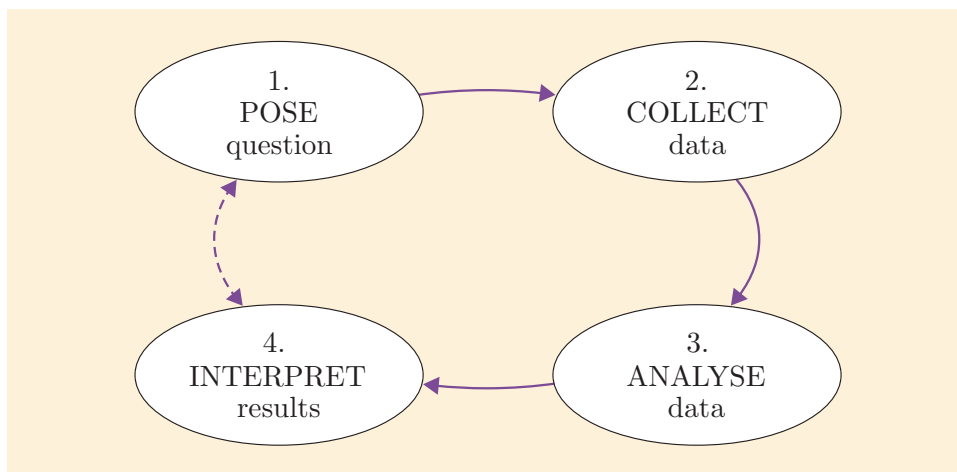


Figure 1 Modelling diagram

This unit concentrates on stage 2, collecting the data. You should by now realise the importance of collecting data that

- can be analysed
- enable you to answer the question under investigation.

Perhaps the most frequent contact that you have with data collection in your everyday life is when you fill in forms or answer questionnaires providing information about yourself, your home, your job, your car or (almost certainly) your OU studies! These can be online or paper and may be for market research companies, government departments or your employers.

Often you are asked to supply the information because you have been selected as one of a relatively small number of people being surveyed, i.e. a **sample**. In other cases, such as the ten-yearly Census in the UK (logos shown in Figure 2), you are part of a large exercise designed to collect information from as many people in the country as it is possible to reach. We shall use the word **census** for any such complete coverage of a population and the word **survey** when a sample is selected from the population.

You may well have wondered, when you are selected to answer questions in a survey, how the answers you give (about your preferences in toothpaste, or the number of children you have) will affect decisions made by whoever commissioned the survey. You may also have considered the question: if your next-door neighbour had been selected instead of you, how much difference would this have made to any decision based on the survey's results? The results of surveys of one kind or another – opinion polls, advertisers' claims – are often in the news; but do they mean anything useful?



Figure 2 The 2011 Census logos for England, Wales and Northern Ireland, and for Scotland



Which was more impressive, the Tower of Suurhusen
or the Tower of Pisa?
If undecided, which way did you lean?

Turning these questions about surveys round and looking at them from the statistician's viewpoint leads to the following question.

Is it possible to gain useful information about a large population (such as all the people in the UK, or all the employees of a large firm) by collecting data about only a relatively small number (i.e. a sample) of them?

The answer, which will be explained in more detail in this unit, is *yes*, provided that the people to be questioned are selected in the correct way.

The **population** need not be a population of people; it could consist of schools, firms, villages, fish, light bulbs, etc. Similar questions can be asked about these populations. For example:

Is it possible to gain useful information about how long light bulbs will last by testing a relatively small number of them?

The answer is again *yes*, provided that the particular items measured or tested are selected in the correct way. Here, though, we shall concentrate on surveys of people.

Section 1 of this unit describes the basic principles of how to select the people to be questioned and introduces a method called *random selection*, or *random sampling*. Section 2 examines the effects of simple random sampling and introduces a modification of this method, called *systematic random sampling*, which is of great practical importance. Section 3 looks more closely at the relationship between samples of the population and the population as a whole. This leads to the idea of a *sampling distribution*, which forms the theoretical basis of methods given in later units for deriving information about the whole of a large population from facts

about a sample taken from it. Section 4 contains an introduction to some further aspects of survey planning. Finally, Section 5 directs you to the Computer Book. You are also guided to the Computer Book at the end of Section 3 as you can choose to work through it from this point if you like.

1 Surveys and sampling

Throughout the previous units, emphasis has been laid on the importance of collecting data that are both relevant to the investigation in hand and reliable. You have also encountered several published sources of data. Now, many of these published sources were based on data that had been collected in surveys. Here is a list of those surveys that have been referred to, with a brief description of them.

1. The survey of prices, carried out each month by a market research company on behalf of the Office for National Statistics; this provides over 100 000 prices used in calculating the Retail Prices Index (**RPI**) and the Consumer Prices Index (**CPI**). (See Section 5 of Unit 2.)
2. The Living Costs and Food Survey (**LCF**), which collects information on the spending pattern of 5000 households. (See Section 5 of Unit 2.)
3. The Annual Survey of Hours and Earnings (**ASHE**) which, each year, collects information on the earnings of about 180 000 people. (See Section 1 of Unit 3.)
4. The Monthly Wages and Salaries survey (**MWSS**) which, each month, collects information about the weekly wages of all employees in about 9000 businesses for use in calculating the Average Weekly Earnings (**AWE**). (See Section 5 of Unit 3.)

All these sources of data have one thing in common: they do *not* collect information about *every* individual member of the population involved (i.e. they are *surveys*, not *censuses*). The whole population of interest is known as the **target population**. Each of these surveys claims to provide reliable information about the *whole* of its target population.

1. For the survey of retail and consumer prices, the exact size of the whole target population is difficult to assess but it is certainly much larger than the 100 000+ prices collected in the survey.
2. The target population of the LCF is all households in the UK. There are about 23 000 000 (23 million) of these.
3. The target population of the ASHE is all employees in the UK. There are about 29 000 000 (29 million) of these.
4. Since the AWE aims to give an overall measure of changes in the wages and salaries of all employees in the UK, the target population is all businesses in the UK. Altogether, there are about 4 800 000 (4.8 million) businesses in the UK. Although businesses employing fewer than 20 people are not sampled, the survey covers approximately half of those in employment in the UK.

The basis for using a survey instead of a census is that, provided the sample is chosen carefully from the target population, the results of the survey can be used to infer the characteristics of the whole target population. We shall see later how this can be done, but first let us consider some of the advantages.

1.1 Why do a survey?



The most common reason for conducting a sample survey rather than a census of the whole population is that the census would be prohibitively expensive in terms of both time and money. For example, if a market research company wished to learn why people prefer to buy *Purr* cat food rather than *Mew* cat food, the expense of questioning everyone in Britain who has a pet cat could not be justified. It may however be practical to survey a sample of 1000 cat owners. Generally the government has greater resources and typically has more important issues to address, but if a survey does provide reliable information about the whole of its target population, then it is certainly much cheaper than collecting this information from every member of the target population. With ASHE, for example, the target population is more than 100 times as big as the sample, so many of the operations involved in collecting the ASHE data would take considerably more money and effort if information about every person in employment in the UK were collected. Some of the operations would not be as much as 100 times as costly, but some would certainly become excessively expensive. Another reason for preferring the survey is that it would take much longer to analyse the larger amount of data from a full census, so the results would be more out-of-date when they were published.

It is certainly true that since only part of the population is included in the sample, the accuracy of the results is threatened, as the characteristics of a sample are very unlikely to be *exactly* those of the whole target population. However, if a suitable method of selection is used in choosing the sample, it is possible to be fairly precise about how large a discrepancy is likely to occur between certain characteristics of the sample and the corresponding characteristics of the target population. The sampling method can then be planned in such a way that the results of the survey are accurate enough for the purpose for which they are needed. Also, in a survey, more care and attention can be given at an individual level than is feasible in a census. This should improve the quality of the data that are gathered, and this will partly offset the uncertainty that arises from sampling.

1.2 Random sampling

In choosing the sample of people to be questioned in a survey, it is important that a suitable method of selection is used. If the statisticians working on the ASHE chose their sample of employees by asking every business in the country how much the managing director earned, for example, then the data collected would not be a very useful measure of the distribution of earnings in the country! A useful sample must be spread evenly over the target population. However, the ASHE statisticians would still not get very accurate information about earnings in the country as a whole by investigating the earnings of a sample of, say, just five people, however carefully they were selected. A useful sample must also be large enough – but how large is large enough? How should a sample be chosen to obtain accurate information about a large population, within constrained budgets?

We require a method of choosing a sample from the target population that is no larger than necessary, because, in general, the smaller the sample, the cheaper the collection of the data. On the other hand, the information collected from the sample must enable us to obtain sufficiently accurate information about the target population; and this means that we cannot choose very small samples. The size of the sample used in a survey has to be a compromise between these two criteria, which can be summarised as **economy** and **accuracy**. Resolving the conflict between these criteria is the aim of a good method of choosing a sample.

The process of carrying out a survey can be briefly described as follows. You start with a target population and from it you select a sample. You then collect data about this sample. From these data, you want to be able to obtain information about the target population. This process is called *inferring back from the sample to the population*. So you want to choose a sample with properties similar to those of the target population.

The ASHE uses the sample of all people whose National Insurance number ends in a particular pair of digits. This is a good method of choosing a sample for the following reason: there is no relationship between people's National Insurance numbers and their earnings, and this implies that the distribution of the earnings of people in this sample is very likely to be similar to the distribution of the earnings of the whole target population. A slightly more precise way of expressing this property of an ideal sample is to say that *a pattern in the sample implies a similar pattern in the target population*. Such a sample is called a **representative sample**.

No method of selecting the members of a sample can be guaranteed *always* to produce a representative sample (unless we select every member of the target population!) but one way of getting close to this ideal is to use a method called *random sampling*. This method will be illustrated by using a very small target population consisting of a fictional household, which contains only four members:

Jim Susan Linda Matthew.

Other factors that affect the cost of a survey will be considered in Section 4.

The UK does not have a system of personal identity cards, and not everyone has a passport. National Insurance numbers and National Health Service numbers are the only two systems that provide almost every adult in the UK with a code number.

Suppose, for the sake of illustration, that we want to investigate the miserliness of this household by asking a sample of individuals from it how mean they are, but that there is only enough money in our survey budget to draw a sample of two people from the household. (Times are hard.)

In this simple situation, we can write down a list of all the possible samples of two different people that we could choose. There are six of them. They are:

- 1 Jim Susan
- 2 Jim Linda
- 3 Jim Matthew
- 4 Susan Linda
- 5 Susan Matthew
- 6 Linda Matthew

As the name ‘random sampling’ suggests, we let chance choose our sample for us. We shall introduce chance into our method of selection by throwing a die.



Figure 3 A pair of dice

First, we must label the six possible samples from the household with the numbers on the six faces of the die: 1, 2, 3, 4, 5, 6. It does not matter which sample gets each label but we shall use the labelling in the list above. Then we can relate the throwing of any one of these numbers on the die to the selection of a particular sample. If we throw a 3, then we select Jim and Matthew.

So long as we do not cheat when throwing the die, and so long as the die is not ‘loaded’ in some way that makes some numbers more likely to come up than others, this method of choosing a sample is an example of **random sampling**, and the sample chosen is a **random sample**. Such a method is also called **random selection**, and we say that the members of the sample are selected, or chosen, at random – or that they are randomly chosen. The characteristic of a random sample is that every possible sample has the same chance of being selected.

This method of random sampling could, in principle at any rate, be extended to larger samples from larger target populations by using a fair (i.e. not ‘loaded’) die with more than six faces. For instance, there are 20 different samples of size three that could be drawn from a household with six members, and we could choose one of these samples by listing them all, numbering them from 1 to 20, and rolling a die with 20 faces (such as that shown in Figure 4).



Figure 4 A 20-sided Roman gaming die from the 2nd Century AD

This might just about be feasible, but things quickly get out of hand with populations and samples of the sort of size that are needed in practice. For instance, suppose you wanted to choose a sample of 100 students from an OU module that has 1000 students in all. The number of possible samples is about 6×10^{139} , and it would clearly be impossible either to write out all the possible samples in a list or to construct a die with 6×10^{139} faces to choose one of them at random. Therefore, we have to develop a slightly different way of choosing our sample of two members of the fictional household out of the population of four. This new way will be much easier to extend to larger samples from larger populations.

The number 6×10^{139} would be written down as a 6 followed by 139 zeros.

What we shall do is to choose the individual people to go into our sample one at a time. Look again at the list of all possible samples.

Jim	Susan
Jim	Linda
Jim	Matthew
Susan	Linda
Susan	Matthew
Linda	Matthew

Each individual appears in the same number (three) of the six possible samples. Therefore, all of the four household members are equally likely to appear in any particular sample that we happen to choose. Let us label the household members, rather than the samples, with numbers:

1 Jim 2 Susan 3 Linda 4 Matthew.

To select the first member of our sample, we throw the die and record the number thrown. Then we select the person who is labelled by this number. (We could use a four-faced die for this if we had one, or we could just use an ordinary six-faced die and ignore any throw which resulted in a 5 or a 6.) To select the second member of our sample, we repeat the above process. However, if the die shows the same number as the first selection, we throw again, because we do not want to include the same person in our sample more than once.

If we require a sample of size two and the numbers thrown were 2 and 3, then Susan and Linda would be selected. If, however, the numbers thrown were 1 and 1, we would ignore the second 1 and throw again. If we obtained the number 4 on the next throw the sample would be Jim and Matthew. Choosing the sample members one at a time like this still has the property that any of the possible samples is just as likely to be chosen as any other, so that conceptually it is no different from the first method we described. It is much more practical to use this one-at-a-time method for larger samples and populations.

In some circumstances, it *is* appropriate to allow samples in which the same individual can appear more than once, though these types of situations are not considered in this unit.

We could choose a sample of three people from a household of size six by numbering the individuals in the household from one to six and throwing a six-sided die at least three times. (More than three throws might be needed to avoid repetitions.) Even for the problem of drawing a sample of 100 students from a population of 1000, the one-at-a-time approach would save having to write out all 6×10^{139} possible samples in a list: we would

You will learn how to use Minitab to generate random numbers in the final section of this unit.

We can also use pairs of digits for target populations of size less than 100, as will be described in Section 2.

just have to write out a list of all 1000 students, number them from one to 1000 and start rolling a 1000-faced die. For a target population of 1 000 000 people we should need a die with 1 000 000 faces!

It may seem impossible to do anything like this! In practice, statisticians use computer programs to generate random numbers which can act in this manner. We shall now see how to use random numbers in this way.

The following random numbers are taken from a set that were generated using Minitab:

9 8 0 6 7 7 4 6 1 6

They can be written as pairs of digits,

98 06 77 46 16 ...,

and are then exactly equivalent to the results of throwing an imaginary fair die with 100 sides labelled 00, 01, 02, ..., up to 99. If you had a target population of size 100, you would probably find it simplest to label the first member 01, the second 02 and so on, with the 99th member labelled 99. Then the 100th member would use the label 00. Then you could use the throws of the imaginary die to select a random sample. As with the real die, if a pair of digits that you have already used in the sample turns up again, you just ignore it and go on to the next pair.

So the pairs of digits at the start of the first row in the list above would select those members of the population labelled 98, 06, 77, 46 and 16. These members therefore form a random sample of size five.

If more than one sample is required from the same target population, then you should not start from the same place in your list of random numbers every time, because this would lead to the selection of the same members of the population in every sample. It is important to start at a *different* point in the list for each sample. The starting point should ideally be selected randomly (using a die or some other procedure). However, to aid explanation, you will usually be told where to start in each case.

Activity 1 Random sample from population of 100

Choose a random sample of size 12 from the population of 100 individuals labelled 00 to 99, using the method described above. A table of random numbers, generated using a computer, is provided as an appendix to this unit. Use successive pairs from row **79** of the random number table, beginning with the first pair in the row, i.e. 52.

You may have found it a little awkward in the last activity to check for repetitions in the sample. In relatively small samples from larger populations than this, repetitions are very rare occurrences in practice.

For a target population of size 1 000 000, we need to use the following labels

00 00 00 00 00 01 00 00 02 ... up to 99 99 97 99 99 98 99 99 99.

Again, the population would probably be labelled 00 00 01, 00 00 02, 00 00 03, ..., up to 99 99 99, 100 00 00, and we should use the random number 00 00 00 for the last member. Then, for the throws of an imaginary die with 1 000 000 sides, we use groups of six digits in the random number table. If we start with the row designated **20**, say, then the first three labels selected will be

59 70 46 36 67 19 12 59 39.

Lottery draws

Major lotteries, such as the UK National Lottery, use special machines to draw the random winning numbers. The draws are open (they are often televised), and the purpose of the machines is partly to put on a spectacle but also to make it transparent that the lottery is fair and the numbers are drawn truly at random. The latter is important as a randomly drawn set of numbers will sometimes look very odd. For example, the six numbers drawn in the UK National Lottery on 11 October 2008 (excluding the 'bonus ball') were all in the twenties – 20, 21, 23, 24, 27 and 28 – despite being a random selection from the numbers 1 to 49.



A UK National Lottery machine

Activity 2 *Random sample from population of 1 000 000*

Choose a random sample of size ten from the target population of size 1 000 000 using labels as described above. Use rows **15** and **16** of the table in the same way as we used row **20** above.

1.3 Properties of simple random sampling

You have learned how to find a random sample of the target population (and been told why it is called a random sample). This process is usually called **simple random sampling** (and the samples chosen are called **simple random samples**) to distinguish it from other random methods, some of which will be described later. The very important *random* nature of the procedure can be more precisely expressed as follows.

In Unit 6, we shall be able to express these properties even more precisely because there we shall encounter probability. This is a measure of chance and it gives us a language for describing random processes.

Simple random sampling

This is a method of selecting a sample in which the possible samples of a given size, n , consist of all possible selections of n different individuals from the population. The sample to be used is chosen in such a way that every possible sample is equally likely to be selected.

One way of doing this is to choose the sample members one at a time in such a way that:

- At each selection, every member of the target population is equally likely to be selected.
- The selection of a particular member of the target population has no effect on the other selections, beyond the requirement that the same individual cannot appear more than once in the sample.

It may seem paradoxical to you that we should be recommending a method of obtaining a representative sample in which chance plays such an important role. One analogy that might help you to see why simple random sampling is sensible is the following.

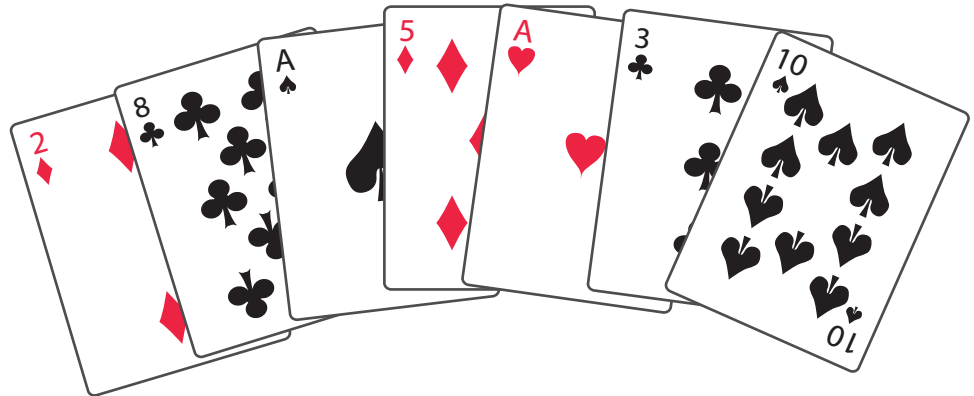


Figure 5 A hand of cards

The process of shuffling a pack of cards well and then dealing a hand is essentially a method of choosing a hand of cards (such as that in Figure 5) by simple random sampling from the pack. If you have played any card game, you will probably be aware that most hands of cards contain a fairly even distribution of suits, and contain a few court cards but not a great many of them. Therefore, they have properties that match the properties of the whole pack, which has an even distribution of suits and just under

25% of the pack is court cards. To put it another way, if you actually wrote down a list of all possible hands of cards, some of them would be unrepresentative in terms of suit distribution or the number of court cards, but most would be representative. Therefore, when one of the possible hands is chosen or dealt at random, it is more likely to be representative than it is to be peculiar.

In the next section we shall look critically at simple random sampling, and see that it is certainly no exception to the statement made earlier: that no method is guaranteed always to produce a representative sample, i.e. a sample from which we can make completely accurate inferences about the population. (Hands of cards consisting entirely of one suit do turn up!) However, randomness is an essential feature of most good methods of choosing a sample.

It is not always necessary, or possible, to use random numbers to choose a random sample. For example, suppose that you wanted to choose a random sample of size ten from a population of 100 fish in a tank. It would probably be very difficult to label each individual fish, and it would be impossible if you wanted to choose a sample of fish from the North Sea.



Figure 6 Netting fish

It would therefore be impossible to use random numbers to choose a sample. Simply selecting ten fish from some caught in a net (Figure 6) is, for many purposes, as good a method as any of choosing this random sample. Unless, for example, you want to measure their size, or how difficult they are to net!

Much of this section has been concerned with general methods. You have seen that a well-chosen sample is an economic and accurate method of collecting data about a population, and that simple random sampling is a good method of choosing a sample. You have seen how to use random numbers to choose a simple random sample from a population with numerical labels. In contrast, the next section will be more specific and more practical. We shall concentrate on a particular target population and choose some random samples from it.

The characteristics of the collection of all possible samples is dealt with more precisely in Section 3.

Exercises on Section 1

Exercise 1 *Random sample from population of 1000*

In this exercise we have a new target population whose size is 1000. Use the random number table in the appendix to choose a random sample of size seven from this population.

Exercise 2 *Random sample from population of 100*

The population in this exercise is of size 100, labelled 00 to 99.

- (a) Choose a random sample of size nine using pairs from row **25**. Start at the third pair, which is 26, and work to the right.
 - (b) Choose a random sample of size 17 using pairs starting at the beginning of row **26**. Move along row **26** to the right-hand end and then go to the next row, designated **27**.
-

2 Random samples

Throughout this section, we shall assume that, just as in Units 2 and 3, we are interested in investigating whether people have been getting better or worse off. To pursue this investigation, we might carry out a survey in which several related, and relevant, questions on this subject are put to a sample of individuals. The questions might be concerned with changes in their income and expenditure, as well as their subjective feelings about their economic well-being.

Our target population will be those people who work in the mythical Sampling Department in a large organisation. These 86 people are listed in Table 1 in alphabetical order of surname. This list is based on a staff list from a real organisation; the names and other details have been changed to preserve confidentiality.

Each person has been given a label. We have also recorded their gender and occupational group. The information in these last two columns will not be used immediately; it will become relevant later, because a person's gender and occupation may have a bearing on how well off he/she is. For choosing a random sample, we need the second column together with some random numbers. We will use the table of random numbers given in the appendix to this unit.

Table 1 Sampling Department staff list (in alphabetical order)

Name	Label	Gender	Occupation*
Alicante-Node, Alphonso	01	M	M
Andrews, Jean	02	F	P
Archer, Simon	03	M	M
Baines, Tom	04	M	P
Baker, Fred	05	M	P
Bates, Sheila	06	F	S
Baxter, John	07	M	P
Best, John	08	M	P
Bidford, David	09	M	P
Bond, Mick	10	M	P
Bramley, Max	11	M	P
Burroughs, Sean	12	M	P
Cameron, Lynne	13	F	P
Carter, Jane	14	F	P
Chapman, Liz	15	F	M
Clark, Rowena	16	F	S
Clarke, Jim	17	M	A
Cluskie, Alex	18	M	P
Cramer, Will	19	M	P
Crofts, Dennis	20	M	P
Crofts, Mary	21	F	A
Crossman, Kim	22	M	S
Daley, Stuart	23	M	P
Damper, Emma	24	F	S
Dev, Mohen	25	M	P
Eisenstein, Bert	26	M	P
Eric, Steve	27	M	P
Estover, Matthew	28	M	P
Fallow, Jim	29	M	P
Flint, Gerald	30	M	P
Foster, Sue	31	F	S
Franks, Abraham	32	M	P
Gowan, Dai	33	M	P
Graham, Bert	34	M	P
Graham, Bill	35	M	P
Grant, Lynne	36	F	P
Gray, Chris	37	M	P
Greenson, Denise	38	F	A
Greenway, Maggie	39	F	P
Hallow, Jean	40	F	A
Hare, Dorothy	41	F	P
Harrison, Sheila	42	F	P
Hewitt, Ray	43	M	P

* P = Professional, A = Administrative, S = Secretarial, M = Manual

Name	Label	Gender	Occupation*
Hopkins, Jane	44	F	A
Howe, Phil	45	M	P
Hutton, Joan	46	F	S
Iron, Donald	47	M	P
James, Patricia	48	F	A
Jolly, Susan	49	F	S
Kapoor, Sashi	50	M	P
Lang, Chris	51	M	P
Light, Phil	52	M	P
Locke, Carol	53	F	S
London, Fred	54	M	P
Lupton, David	55	M	P
McCarthy, Keith	56	M	P
McCraig, Frank	57	M	P
Masterton, Dick	58	M	P
Menton, Christine	59	F	S
Menton, Pete	60	M	P
Munn, Sharon	61	F	P
Neilsen, Rob	62	M	P
Osterley, Rebecca	63	F	S
Patel, Deepak	64	M	P
Pinder, Andrew	65	M	P
Redman, Guy	66	M	P
Redstar, Pamela	67	F	S
Ricardo, Dan	68	M	P
Roberts, Christine	69	F	S
Rowan, George	70	M	P
Sandford, Dave	71	M	P
Shah, Anjali	72	F	S
Singh, Meera	73	F	S
Stratford, Peter	74	M	P
Thompson, Anna	75	F	S
Thompson, Jack	76	M	P
Trumpington, Pat	77	F	S
Truscott, Karen	78	F	S
Turner, Richard	79	M	P
Tyndale, Babs	80	F	S
Watson, Eleanor	81	F	P
Wilton, Larrie	82	F	P
Winston, Chuck	83	M	P
Woodhouse, Paul	84	M	M
Wu, C. C.	85	F	M
Yeo, Tara	86	F	A

* P = Professional, A = Administrative, S = Secretarial, M = Manual

2.1 Choosing some samples

In Subsection 1.2 we described a way of using random numbers to choose a sample from a target population of size 100. A small adaptation of this method will enable you to choose a sample from the target population of size 86. In the department list (Table 1) the members of the target population are labelled 01, 02, 03, $\dots$, and so on, up to 84, 85, 86. You could therefore use pairs of digits to select members of a sample just as you did for the 100 labels in Subsection 1.2 but, trying this method, if you randomly selected 93 as your starting pair of digits you would be unable to select a person with this label. You should simply ignore this pair and go on to the next pair in your list of random numbers.

To use pairs of digits as throws of an 86-sided die: simply ignore any pair of digits that is not one of the 86 labels in the list of the target population.

Example 1 *Random sample from population of 86*

We shall now use row **53** of the table in the appendix to choose a sample of size ten from our target population. We work along the pairs of digits in this row until we have ten labels in the range 01 to 86, ignoring all pairs of digits outside this range.

Row 53	93	46	82	67	64	48	91	74	85	94	40	51	30
Label of selected individual	X	46	82	67	64	48	X	74	85	X	40	51	30

Looking for these labels in the department list we find the sample listed in Table 2. This table shows the name and label of the ten people selected for the sample and also their gender and occupation. The last column, which is headed ‘Response’, is explained below.

Table 2 A sample of ten staff

Name	Label	Gender	Occupation	Response
Hutton, Joan	46	F	S	No
Wilton, Larrie	82	F	P	Yes
Redstar, Pamela	67	F	S	Yes
Patel, Deepak	64	M	P	No
James, Patricia	48	F	A	Yes
Stratford, Peter	74	M	P	No
Wu, C. C.	85	F	M	Yes
Hallow, Jean	40	F	A	Yes
Lang, Chris	51	M	P	No
Flint, Gerald	30	M	P	No



Example 1 is the subject of Screencast 1 for Unit 4 (see the M140 website).

Now that we have selected a random sample of people in the department, we can use it to investigate whether people think they are getting better off. We might start by asking the ten people a straight question, ‘Do you feel that you are better off now than you were twelve months ago?’ and ask for a straight ‘Yes’ or ‘No’ response. Suppose that the answers given to this question are those shown in the last column of Table 2.

We shall discuss the choice of question a little more in Subsection 3.1.

In the sample, there were five ‘Yes’ responses and five ‘No’ responses. Can we say that there would be equal numbers of ‘Yes’ and ‘No’ responses in the whole population? In other words, how representative is the sample of the target population? Is there anything we can do to check its representativeness? We cannot check whether the responses to the question are representative because we do not know the responses of the whole target population. However, we can use the information in the columns headed ‘Gender’ and ‘Occupation’ in Table 1 to check how representative the sample is for these characteristics. If the sample is unrepresentative in terms of gender or occupation, it is less likely to be representative in terms of whether people feel they are getting better off. However, before we can do this check, we must analyse the information contained in these columns. The information contained in Table 1 about the structure of the target population is summarised in Table 3, which lists the number of department staff of each gender and the number in each occupational group; there are eight different gender–occupation categories in all.

Table 3 Department staff analysed by gender and occupation

	Male	Female	Total
Professional	46	10	56
Administrative	1	6	7
Secretarial	1	17	18
Manual	3	2	5
Total	51	35	86

Since the staff list is based on that of a real organisation, it reflects the fact that in many British organisations the gender balance in different occupations remains uneven. Out of 56 people in the professional group, 46 (82%) are male, whereas 17 out of the 18 secretarial staff (94%) are female. The module team chose to use this particular example not because we approve of the status quo on gender balance, but because we want to demonstrate the important role that statistics can play in investigating such issues and monitoring change.

Table 3 can be used to compare the target population with any sample from it and thus to check on whether the sample is *representative* with respect to gender and occupation. To do this, it is usually better to express the number in each category as a percentage of the total: 86. This has been done in Table 4.

Table 4 Percentages of department staff by gender and occupation

	Male	Female	Total
Professional	53.5	11.6	65.1
Administrative	1.2	7.0	8.1
Secretarial	1.2	19.8	20.9
Manual	3.5	2.3	5.8
Total	59.3	40.7	100.0

Note that all the percentages in Table 4 were found by dividing the corresponding entry in Table 3 by 86, multiplying by 100 and then rounding to one decimal place. Therefore, some of the figures in the ‘Total’ row and column of Table 4 do not correspond exactly to the totals of the rounded values in the table, because of the small inaccuracies introduced by rounding.

Using this information we can now demonstrate that the sample in Example 1 is not very representative. Two facts will suffice.

- The majority of the sample – six out of ten, or 60% – consists of women, compared with only 40% of the population.
- 20% of the sample are in the administrative category, and 50% are in the professional category, compared with the proportions of about 8% and 65%, respectively, in the population.

This sample should thus be described as unrepresentative with respect to gender and occupation. It would not be possible to reproduce all the percentages in Table 4 exactly in a sample of only 10, of course, but you might hope to get rather closer than we did in this sample. The sample was chosen by random sampling but it has turned out to be unrepresentative of the population in terms of gender and occupation. Therefore, if you were able to do a similar comparison for responses to the question about how well off people felt, you might well find that the results from the sample did not agree with those of the population.

As you would expect intuitively, all other things being equal, the larger the sample chosen from the population, the more representative it is likely to be, and the closer the characteristics of the sample will be to those of the population.

Activity 3 *Sampling from the Sampling Department*

Choose a random sample of size 20 from the department list using the random number table provided in the appendix to this unit starting at the beginning of row **2**.

Note the gender and occupation of each individual selected and then comment on the representativeness of the sample with respect to gender and occupation.

2.2 Systematic random sampling

You should now be able to appreciate how time-consuming and tedious it would be to choose even a moderately large sample from a fairly large population using simple random sampling. The sizes of the samples we have chosen so far are trivial compared to the sampling requirements of some official, academic and market research investigations.

An alternative method, which provides a quicker and easier means of choosing a sample from a list of the target population, is **systematic random sampling**. This method is similar to that used to choose the sample for the ASHE (Annual Survey of Hours and Earnings), which selects one in 100 of the National Insurance numbers (which are themselves issued sequentially). The ASHE does not select these labels randomly but selects all the labels with the same pair of final digits. The only randomness in this procedure comes in choosing which one of the 100 pairs of digits to use. Having made this choice, the selection is completely systematic and can be described as selecting every 100th label in the ordered list of labels.

So, as the National Insurance numbers are just labels, we can use the labels 01 to 86 of our population in Table 1 in a similar way.

In practice, for a real survey the sample would be drawn using a computer. Computers do not find jobs tedious (or enjoyable!). In Section 5 you will learn to use Minitab to draw random samples.

Example 2 *Sampling every eighth individual*

Using a similar procedure to that above, select a sample of about one-eighth of our target population using the labelled list as follows.

Step 1 Decide where to start by randomly choosing a label from the first eight labels, 01 to 08. This label is the **random start**. Suppose that it is 04.

Step 2 Select the remaining individuals from the population by systematically selecting every eighth label. The number eight is the **sampling interval**.

This gives the following 11 labels.

04 12 20 28 36 44 52 60 68 76 84

So the sample with sampling interval eight and random start 04 is as shown in Table 5.

Table 5 A sample of every eighth individual

Name	Label	Gender	Occupation
Baines, Tom	04	M	P
Burroughs, Sean	12	M	P
Crofts, Dennis	20	M	P
Estover, Matthew	28	M	P
Grant, Lynne	36	F	P
Hopkins, Jane	44	F	A
Light, Phil	52	M	P
Menton, Pete	60	M	P
Ricardo, Dan	68	M	P
Thompson, Jack	76	M	P
Woodhouse, Paul	84	M	M



Example 2 is the subject of Screencast 2 for Unit 4 (see the M140 website).

In the sample selected in Example 2 there are nine professionals, one administrator, one manual worker and no secretarial staff. Also, there are nine men and only two women, compared to a ratio in the whole population of six to four. Overall, the sample is not very representative of the whole target population.

This shows that a systematic random sample *need* not be any more representative than a simple random sample. However, there are two main reasons for using systematic random sampling: one is to save time, and the other is that in certain special circumstances (which we shall come to later) systematic sampling *does* tend to produce more representative samples.

This method does not always give samples of exactly the same size. This is illustrated in the following example.

Example 3 A second systematic sample

Suppose the random start is 07 and we select every eighth label (i.e. we use the same sampling interval eight). Then we get only these ten labels:

07 15 23 31 39 47 55 63 71 79.

In practice, these discrepancies in size hardly ever matter, as the sample size will only vary by one, and typical sample sizes in actual samples are usually several thousand.

This second sample, with sampling interval eight and random start 07, is shown in Table 6.

Table 6 A second systematic sample

Name	Label	Gender	Occupation
Baxter, John	07	M	P
Chapman, Liz	15	F	M
Daley, Stuart	23	M	P
Foster, Sue	31	F	S
Greenway, Maggie	39	F	P
Iron, Donald	47	M	P
Lupton, David	55	M	P
Osterley, Rebecca	63	F	S
Sandford, Dave	71	M	P
Turner, Richard	79	M	P

In this sample there are seven professionals, one manual worker, two secretarial staff and no administrators. The ratio of men to women is almost exactly that of the whole population. So this happens to be a more representative sample than the previous ones as regards gender and occupation.

Activity 4 A systematic sample of one-seventeenth

Select a systematic random sample of about one-seventeenth of the department. To find the random start, take the first pair of digits in the range 01 to 17 from row **3** of the random number table in the appendix to this unit. Analyse the sample with respect to gender and occupation, and comment on how representative it is in these respects.

Activity 5 A systematic sample of one-quarter

Choose a systematic random sample of about a quarter of the department. This time, take the first digit in row **29** in the range 1 to 4 as your random start. Analyse the sample with respect to gender and occupation and comment on how representative it is in these respects.



From the last two activities, and the examples of simple random sampling in Subsection 2.1, you should now be able to appreciate that systematic random sampling is much quicker to do 'by hand' than simple random sampling, but that it does not necessarily provide samples which are more representative of the target population.

In some circumstances systematic random sampling will do no better and no worse, on average, than simple random sampling in producing representative samples. However, in other circumstances it might do much worse: for example, suppose that you have a list of people in which each consecutive pair are a married couple with the husband always appearing first and the wife second. If you take a systematic random sample from such a list and the sampling interval is an even number, then the sample

will consist entirely of men or entirely of women, depending on whether the random start is an odd or an even number. This shows that care is needed in the use of systematic random sampling: it is hazardous whenever the list of the population contains such regularities. A case as extreme as this could easily be recognised, but if the regularity is less distinct, and hence not noticed, then the problem is more serious.

There are circumstances, though, in which systematic sampling is likely to do *better* than simple random sampling. Suppose that the department list in Table 1 had been ordered by occupation and gender instead of simply being in alphabetical order of names. That is, suppose that all the female professionals were listed first, followed by all the male professionals, then all the female administrators, then all the male administrators and so on. Imagine drawing a systematic sample of a quarter of the department from a list in that order. The sample would inevitably include about a quarter of the female professionals, a quarter of the male professionals, a quarter of the female administrators – in fact, about a quarter of each gender–occupation group. It would therefore be very representative.

This is a kind of *stratified sampling*, a concept you will learn more about in Section 4.

In simple random sampling, all possible samples are equally likely to be chosen. The method tends to work well because most but not all of the possible samples are reasonably representative. In systematic sampling, the number of different samples it is possible to obtain is much smaller. There are only four possible systematic random samples of a quarter of the population in Table 1, because there are only four possible values for the random start. (By contrast, a simple random sample of 21 people from the same population, about a quarter of the population, would be one chosen at random from about 6×10^{19} possible samples.) If the population were listed in gender–occupation order, then all four possible systematic random samples would be representative, so that systematic sampling is bound to do well. However, in a situation like the list of married couples, all possible systematic samples would be unrepresentative, so that systematic sampling is bound to do badly. In many circumstances, though, the population will be listed in some order that has nothing to do with the features of the population it is important to represent; then systematic random sampling is likely to be no more and no less representative than simple random sampling.

To summarise, we have the following properties of systematic random sampling.

Systematic random sampling

Systematic random sampling is easier to carry out than simple random sampling and is very often used for choosing samples from large populations.

- It can produce very unrepresentative samples if the list of the target population is structured in certain ways.
- It produces random samples that are at least as representative as those produced by simple random sampling, provided the target population is listed in a suitable way.
- In certain cases, systematic random samples are considerably more representative than simple random samples.

In this section you have learned how to choose both simple and systematic random samples, using a labelled list of the target population, and you have learned about some of the properties of the two methods.

Exercises on Section 2

Exercise 3 *Selecting more simple and systematic samples*

This exercise is on choosing both simple and systematic random samples. After choosing each of the following samples from the list in Table 1, draw up a table similar to Table 3 (in Subsection 2.1) to analyse the sample by gender and occupation.

- Choose a simple random sample of size eight using row **5** starting at the beginning.
- Choose a simple random sample of size 12 using row **10** starting at the beginning.
- Choose a systematic random sample with sampling interval nine and random start 05.
- Choose a systematic random sample with sampling interval ten and random start 08.

3 Patterns in the samples

So far in this unit we have looked at individual samples from a target population and considered whether a sample is representative of its target population. In the last section, some of the samples we drew did seem to be representative of the target population; others did not. In this section we shall take a different view of sampling. We shall consider all the possible samples of a given size that could arise when choosing a sample from a given population. You will see that patterns arise in such collections of all possible samples, and that these patterns provide information about the representativeness of samples. Here, we shall look at samples from one particular population, but similar methods can be used to describe patterns in collections of samples from any population.

3.1 Population values and sample values

In Section 2, the aim of sampling from the population was to investigate whether people were getting better or worse off. (That was why we wanted a sample that was representative in terms of gender and occupation – factors likely to determine how well off someone is.) Here, we shall continue with the same aspect of this investigation: determining people’s subjective feelings about changes in their own economic circumstances.

There are several methods of obtaining such information, but, because of its subjective nature, they nearly all involve asking people questions. Therefore, a reasonably good method of obtaining the required information is to question a relatively small sample of the target population. The most straightforward question we could ask on this topic is a question such as the following.

Are you better off than you were twelve months ago?

However, such a blunt question would probably not produce very useful data. There are many reasons for this, but one of the most crucial is that different people will interpret it in different ways. (To test this claim, try asking your friends this question and note the way in which they interpret it.) A better question for our investigation is as follows.

Considering what has happened to your earnings, the way prices have changed and changes in other circumstances, do you feel that you are now better or worse off than you were twelve months ago?

This question still leaves one problem that always occurs when investigating people’s subjective feelings. If someone asked you a question like this, you might well reply at length describing your personal circumstances and events during the year. Such responses are hard to analyse, so it is very common to ask the respondent to classify his or her answer into one of a small number of categories.

This is most commonly done through a **Likert scale**, named after Rensis Likert (1903–1981), whose work underlies its popularity.



Rensis Likert (1903–1981)

A Likert scale has a number of ordered categories, and respondents tick one of them to specify their level of agreement or disagreement with a statement. For the above question, the following request could be added.

Please tick the phrase that best describes your feelings.

Much better off
Somewhat better off
About the same
Somewhat worse off
Much worse off

☐
☐
☐
☐
☐

This makes it much easier to compare one person's answer with another's and to summarise people's answers. Analysis of the answers is yet further simplified if each response is expressed as a number from 1 to 5 as follows.

Much better off	5
Somewhat better off	4
About the same	3
Somewhat worse off	2
Much worse off	1

So, the better off a person feels they have become, the higher the number we use to label their response.

It is important to realise that the numbers are being used here simply as labels that come in a helpful order. There is no implication that, for instance, 'Somewhat better off' is twice as good as 'Somewhat worse off', just because 4 is twice 2. In fact, the labels for the responses could have been chosen as a, b, c, d, e, rather than 1 to 5.

There are snags. The simplification obscures the individual details of what people might have said if they had been given the opportunity, and you might still worry about whether one person's 'somewhat better off' is the same as another's.



If we choose a sample of people from the target population and ask them this question, then we shall know what those people's answers are: these are the **sample data**. We shall then wish to infer from these sample data information about how the whole of the target population would have answered this question had we asked them all. More precisely, the **response** to the above question is 1, 2, 3, 4 or 5, and we shall want to infer back from the **sample values** of this response to values of this response for members of the target population as a whole. These values for the whole target population are the **population values** of the response.

3.2 All possible samples

The examples in Section 2 demonstrated that *any* method of choosing the relatively small sample required can produce a sample that is not very representative of the target population. Although the best methods of choosing a sample are designed to produce representative samples as consistently as possible, none of them guarantees to do so without fail. However, the samples we analysed in Section 2 suggest that, for all but the smallest sample sizes, either of the random sampling methods (simple or systematic) is likely to produce a sample that is sufficiently representative to justify inferring back to the population from facts about the sample.

The samples in that section also suggested that if you choose a larger sample, then you are more likely to choose a representative sample. The reason for this is that although the results from an individual, randomly-chosen sample may well have no clear pattern, the results obtained from the collection of *all possible* samples of a fixed size has a very distinctive pattern for all but the smallest sample sizes.

We will examine some of these patterns. To do this, it is necessary to imagine that we know *all* the relevant information about the target population. We can then consider what samples taken from that target population might look like. That is, imagine that a census was carried out in which every individual in the target population was asked the question we are interested in, and that we knew what all the responses were. In the rest of this section we shall take this convenient, though rather unrealistic, omniscient view.

Imagine first that the target population is 1000 individuals whose responses to the question (i.e. the population values of the response) are already known to be as described in Table 7.

Table 7 Population values of the response

Response	Rating	Number
Much worse off	1	300
Somewhat worse off	2	100
About the same	3	200
Somewhat better off	4	300
Much better off	5	100
Total		1000

We are now interested in the responses of *all* the possible samples of a fixed size that could be obtained from this population by simple random sampling. Even for fairly small sample sizes the numbers involved at this stage are quite large. There are 499 500 possible simple random samples of size two, 166 167 000 of size three, 41 417 124 750 of size four, and so on.

With such large numbers of samples to consider, it may seem impossible to deduce anything at all sensible about these collections of all possible samples. This problem is made easier because, very often, our main interest lies in just one, or a few, properties of the sample and the population. Suppose, for instance, that we are particularly interested in the median of the responses for the population, perhaps because we want a measure of location for the population's responses.

Activity 6 *Population median for Likert data*

Find the median of the responses of the population described in Table 7.

The median calculated in Activity 6 is often called the **median of the response over the whole population** (or, more briefly, the **population median response**, because the median of a population is often called the **population median**). It was possible to find the population median response, in the way you have just done, only because we have imagined that we know all the population values of the response. In a practical situation, you would have data from only a sample from the population. You could calculate the median of the responses in the sample, of course, but what would that tell you about the population median response? To answer this question, we need to consider patterns in the medians in the collection of all possible samples.

Many useful methods have been devised to find and describe the patterns in the collection of all possible samples of a fixed size. These methods typically identify properties of interest (such as the property 'median is 3') and then, for each property, calculate the proportion of samples in the collection that have that property. The results from applying one such method will be illustrated in the next subsection, using the target population described in Table 7.

Calculations underlying the method use the rules of probability, which will be introduced in Unit 6.

3.3 Pictures of patterns

Suppose that we choose a very small sample, of size three, from our target population of size 1000. There are 166 167 000 possible samples of size three.

Although not impossible, it would be quite complex to picture the responses of all three individuals in each of these millions of possible samples of size three. It is more straightforward to picture the millions of medians of these sample responses. We can then look for patterns in this batch of medians.

Activity 7 Median responses in samples of size 3

Table 8 shows the responses of six typical samples (A to F) of size three from the target population. So, for example, in Sample A the first person who was asked replied ‘somewhat better off’ and so the result was labelled 4, the second person’s response was labelled 1, and the third person’s was labelled 2. The median of these three responses is found by rewriting them in numerical order, 1, 2, 4, and then finding the middle value, which is 2.

Write down the median of each of the six batches of sample responses.

Table 8 Responses of the people in six samples of size three

Sample	1st person	2nd person	3rd person
A	4	1	2
B	5	4	1
C	1	4	4
D	1	5	3
E	1	1	1
F	3	5	5

As you have probably realised from this activity, the median of the responses of a sample of size three from this population is either 1, 2, 3, 4 or 5. We shall call such a median a **median response**. It is possible, therefore, to describe the medians of the responses of *all* the 166 167 000 samples of size three by stating how many of them are 1, how many are 2 and how many are 3, 4 and 5. These numbers can be calculated using the rules of probability, and their approximate values are given in Table 9 (where, for example, ‘359 hundred thousand’ means 35 900 000).

Table 9 Median responses of all samples of size three

Median response	1	2	3	4	5
Approximate number of samples (hundred thousands)	359	226	492	539	46

In Table 10 these numbers are expressed as *proportions* of the total number (166 167 000) of samples of size three. This will enable us to look at the pattern, if any, in these sample median responses and to compare the pattern in these medians with the patterns obtained in the same way from samples of other sizes.

Table 10 Median responses of all samples of size three

Median response	1	2	3	4	5
Approximate proportion of samples	0.216	0.136	0.296	0.324	0.028

(These proportions are obtained by dividing the entries in Table 9 by 166 167 000.)

We have displayed these proportions graphically in Figure 7(a), which is a picture of a **sampling distribution**. It is the **distribution of the median response of the sample**; this is often shortened to the

Proportions, such as those used in Table 10, are a very common way of describing such large collections of numbers.

distribution of the sample median (because the median of a sample is often called the **sample median**).

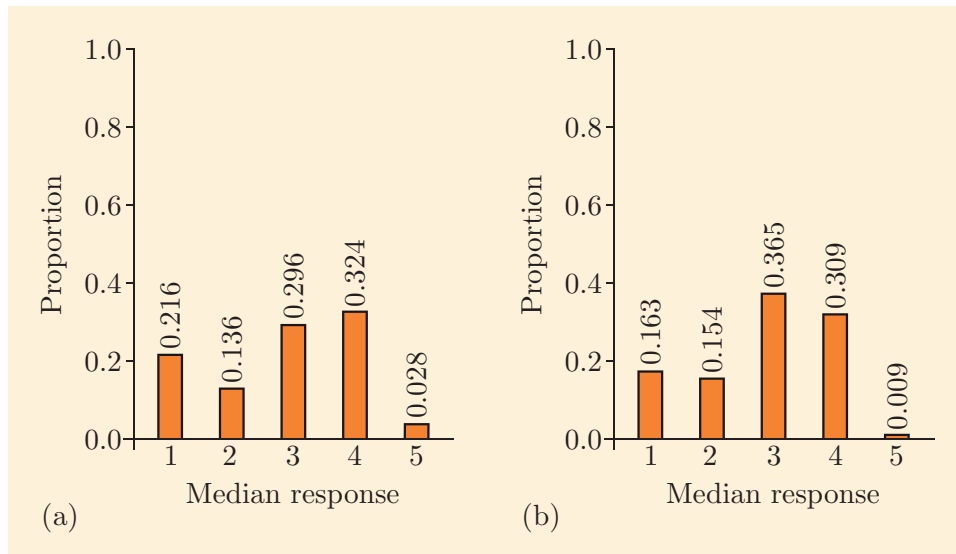


Figure 7 (a) Approximate proportion of samples of size three with each median response; (b) Approximate proportion of samples of size five with each median response

The pattern in Figure 7(a) is not very clear-cut. Not many of the samples have median 5; but one cannot say much more than that. In Activity 6 you found that the median response for the population as a whole was 3. Nearly one-third of the samples also had median 3 – but even more of them had median 4, and large numbers had median 1 or 2 as well. In Section 2 we found that larger samples tended to be more representative of the population. Is this true in terms of medians?

To investigate this, it is useful to have a similar description and picture of the median responses of all the samples of size five (and larger sample sizes). The picture corresponding to Figure 7(a) for the eight trillion (8 000 000 000 000) or so median responses of each of the samples of size five is shown in Figure 7(b).

The proportions here describe the distribution of the sample median for samples of size five. It tells us that about 0.163 of the samples of size five (i.e. 16.3%, or rather more than 1.3 trillion samples) have median response 1, about 0.154 of them have median response 2, about 0.365 of them have median response 3, about 0.309 of them have median response 4 and only about 0.009 of them have median response 5. This is another sampling distribution and it enables us to summarise very concisely all eight trillion samples of size five. Furthermore, it is precisely the type of summary picture we need to compare different sample sizes.

Comparing Figures 7(a) and 7(b), you can see, for instance, that a greater proportion of the samples of size five have a median of 3 (the population median response) than was the case for the samples of size three. How does the picture change as the sample size increases further?



You have now covered the material related to Screencast 3 for Unit 4 (see the M140 website).

3.4 Different sample sizes

Figure 8 contains pictures (corresponding to Figures 7(a) and 7(b) in Subsection 3.3) of the distributions of the sample median for several different sample sizes. For each sample size n there are a huge number of possible samples, each of which has a median, and the picture for sample size n shows the proportion of those medians which are 1, the proportion which are 2, the proportion which are 3, and so on.

Here, we use median as shorthand for median response.

Activity 8 *Effect of sample size*

Describe the most obvious change in the distributions in Figure 8 as the sample size n gets larger.

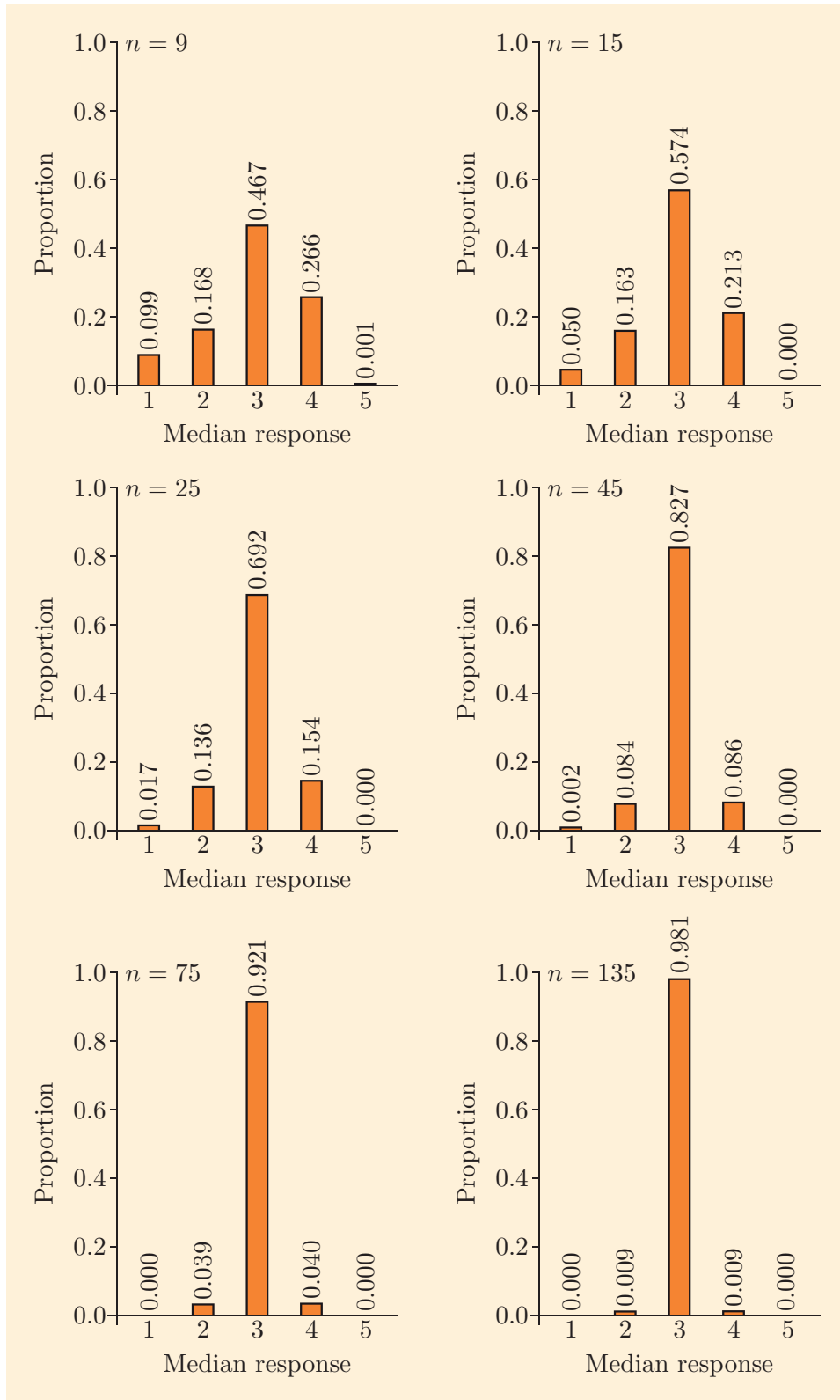


Figure 8 Approximate proportion of samples with each median response for various sample sizes

We have found that as the sample size increases the sample median becomes much more predictable and is much more likely to equal 3, which is the value of the population median. One important consequence of this is relevant to any investigation using samples, including those we considered in Sections 1 and 2.

If you choose a simple random sample of size five from the population described in Table 7, then you are, for example, more likely to choose one with median 3 than you are to choose one with median 5. This is because if you use simple random sampling, then each sample is equally likely to be chosen. You are therefore much more likely to choose one of the large number of samples with median 3 than one of the relatively much smaller number of samples with median 5.

If you choose a larger simple random sample, of size 15 say, then you are more likely to choose one with median 3 than you are to choose one with median not equal to 3; and if you choose a simple random sample of size 135, you are almost certain to choose one with median 3. Now there are an enormous number of possible samples of size 135 and *before* you choose one at random you have no idea which one will be chosen. However, you *can* nevertheless predict with reasonable confidence that its median will be 3. The larger the size of your random sample, the more certainly you can predict what its median will be.

The patterns in Figures 7 and 8 can be described in words as follows. For all but the smallest sample sizes, the sample medians show a very clear and precise pattern: they are nearly all 3. As you found in Activity 6, the population median response is 3. Therefore, as the sample size gets larger, it becomes more and more likely that the sample median response will be the same as the population median response. In this precise sense, the pictures show that larger samples are more representative.

This type of pattern is very common. In general, patterns in sampling distributions from samples of different sizes show that larger samples are more representative. There is also usually a connection between patterns in the population values and patterns in collections of samples from that population (i.e. patterns in sampling distributions).

If, as here, we know the population values, then we can picture their distribution and thus see the patterns in them. The distribution could be pictured on a stemplot for small populations, but for a population of size 1000 this is not a very convenient picture. A common alternative is to use pictures like those used for the sampling distributions in Figures 7 and 8. As with the sampling distributions, we express each number in Table 7 as a proportion of 1000, the population size, and list these proportions on the picture. Thus Figure 9 is a picture of a **population distribution**. We shall study further pictures of population distributions in later units.

The number of samples of size 135 is about 3×10^{170} : written out, this would be 3 followed by 170 zeros.

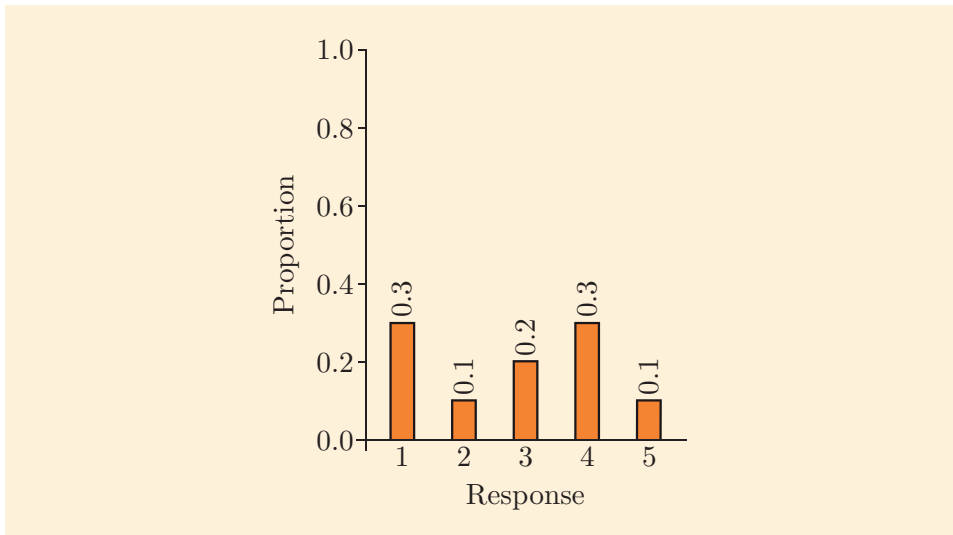


Figure 9 Proportion of members of population with each response

In statistics, interest often focuses on patterns that arise in the collection of all samples of a fixed size. These patterns lie behind many of the methods of analysing sample data that you will meet in later units. In the example we have been discussing, the patterns allowed us to say how likely it is that the sample median response is equal to the population median response. They could also tell us how close the sample median response is likely to be to the population median response; for example, for a sample size of 25 or above, the sample median response might be 2 or 4 (one away from the median) but is very unlikely to be 1 or 5 (two away). More generally, such patterns allow us to say how likely it is that a random sample will be representative in a particular sense, and they allow us to quantify how unrepresentative it is likely to be.

It is important to appreciate that these patterns *can* be described. This is done using sampling distributions. Pictures like those in Figures 7 and 8 are used to summarise sampling distributions and hence show patterns. They are also very useful for describing population distributions (as in Figure 9). So here are some activities based on the pictures in Figures 7 and 8.

Activity 9 *Most likely sample median*

For samples of size three (Figure 7), which value has the largest proportion of the median responses (i.e. what is the most likely median of a simple random sample of size three)?

Activity 10 *Sample median equals population median?*

For which of the sample sizes covered by these pictures (Figures 7 and 8) is it true that over 60% of the samples have median 3?

Another use for patterns of this kind is in choosing the sample size for a survey. Suppose that, for some reason, you were particularly interested in finding out the median of this population, on the basis of sample data. You could do this by finding the sample median and using it as an *estimate* of the population median. The patterns in Figures 7 and 8 show that this estimate would be fairly likely to be wrong if the sample size was only 3 or 5, but almost certain to be right if the sample size was 75 or 135. Such considerations would allow you to choose an appropriate sample size.

In this module there is not time to explain any further how to decide the size of sample which is needed for a particular survey, but one important point is that this does not depend greatly on the size of the target population. Figure 8 demonstrates that a sample of size 75 is very likely to lead to an accurate estimate of the median of a population of 1000 individuals whose responses follow the pattern shown in Figure 9. If the general pattern of responses for the population of the whole of the UK were similar to that shown in Figure 9, then a sample of size 75 would also be very likely to lead to an accurate estimate of the median response for the UK population, even though the UK population consists of well over 60 million individuals rather than 1000.

The most important general points that have been covered in this section are that the collection of all possible samples of a given size has a pattern, that some aspects of this pattern are very precise for all but the smallest sample sizes, and that in looking for such patterns it can be very useful to describe and picture distributions by expressing them in terms of proportions. The last two sections of this unit return to some practical matters involved in planning and running surveys.



You have now covered the material needed for Subsection 4.1 of the Computer Book.

Exercises on Section 3

Exercise 4 *Proportions for a sample of 9*

For sample size 9 (Figure 8, Subsection 3.4),

- (a) approximately what proportion of the samples have median 1?
 - (b) approximately what proportion have median 2?
 - (c) approximately what proportion have median less than 3?
 - (d) approximately what proportion have median greater than 3?
-

Exercise 5 *A different population*

Suppose a different population of 1000 people gave the following responses:

Response	Rating	Number
Much worse off	1	200
Somewhat worse off	2	400
About the same	3	200
Somewhat better off	4	100
Much better off	5	100
Total		1000

- (a) What is the median response for this population?
- (b) Figures A, B and C show three distributions of a sample median. One is for a sample of size seven from the above population, one is for a sample of size 21 from the above population, and one is for a sample of size 21 from a different population. Giving your reasons, say which figure relates to which sample.

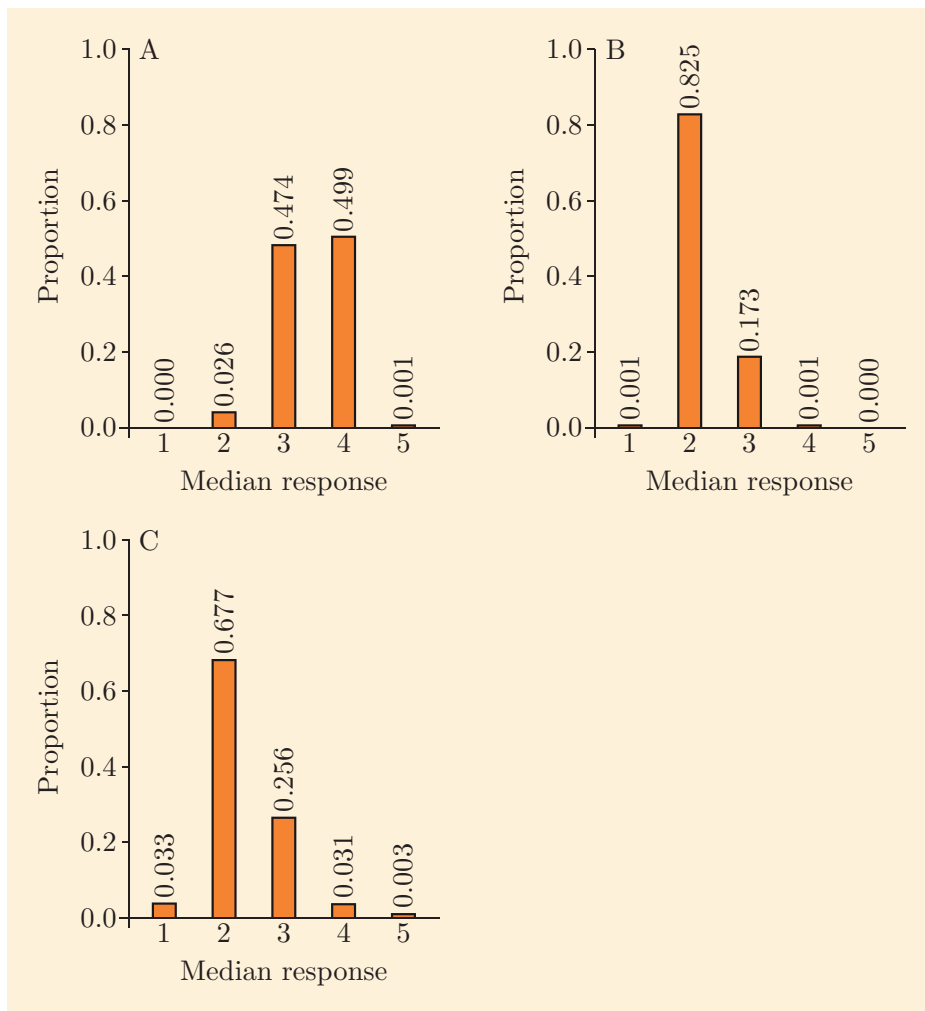


Figure 10 Distributions of three sample medians

4 More sampling methods

Section 2 of this unit introduced two ways of selecting a random sample for a survey – simple random sampling and systematic random sampling. In Section 4, more ways of choosing a sample for a survey will be introduced: stratified sampling, cluster sampling and quota sampling. Before this, in Subsection 4.1, you will learn about types of error that are associated with results obtained using survey data.

4.1 Types of error

If you intend to survey a population by investigating a random sample and inferring from data about this sample back to the population, then it is most unlikely that the results you get from the sample will be identical to those you would have got if you had obtained results from every individual in the population. For example, if you were interested in the mean, the mean of the sample will almost certainly not be the same as the mean of the population, although you hope that the two will not be very different. Statisticians refer to this difference as an **error** and there are several different types of error.

First, there is what is known as **sampling error**. As we saw in Section 3, different samples contain different individuals, and although there is a pattern in the possible results, we cannot know where our particular sample lies in the pattern. So there is variability due to sampling. This is the source of sampling error.

Second, there may be error introduced by using a poor sampling scheme. An example of this is a mobile phone survey where the sample is selected from a listing of mobile phone numbers. Selected people are contacted by phone. This survey has a **bias** in that people who do not own a mobile phone or who have chosen not to have their number listed could not possibly be included in a sample. A survey based on the electoral register would also include a bias against people who move house frequently. Another situation in which bias arises is *quota sampling*, which will be described in Subsection 4.5.

Third, there are other **non-sampling errors** which can arise from a variety of causes; for example, errors in recording responses or in transferring them to a computer, failure to contact individuals who are supposed to be included in a sample or refusal of people to cooperate with the interviewer.

Both the second and third types of error can be reduced or eliminated by planning the survey properly, by employing experienced interviewers and by careful checking. It is impossible to eliminate the first type, the sampling error, because this is inherent in the process of sampling. However, design of the survey can reduce the sampling error, as we shall see in this section.



Other things being equal, a larger sample size gives more accurate results but also leads to higher costs. In an ideal world with no resource constraints, sampling error could be eliminated completely by investigating the whole target population. However, in the real world the costs of collecting reliable data are considerable, so survey planning must involve careful consideration of the resources available.

The Gallup Poll and George Horace Gallup

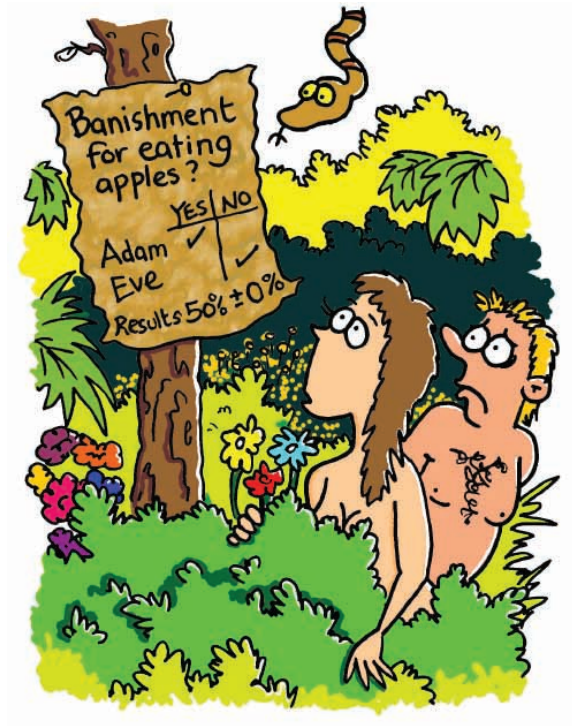
George Gallup (1901–1984) made important advances in survey sampling methods and founded his own polling company (which became the *Gallup Organization*) in 1935. The company came to prominence the following year when it used a survey of 50 000 respondents to correctly forecast that Franklin Roosevelt would defeat Alf Landon for the U.S. presidency. An influential magazine at the time, the *Literary Digest*, conducted a much larger survey but incorrectly predicted that Landon would win. Moreover, Gallup's company correctly forecast the prediction that the *Literary Digest* would make, by following the sampling procedure they used, though with a much smaller sample size. The *Literary Digest* had sampled a list of its own subscribers and lists of car owners and telephone users, so that (in 1936) it was only sampling from the more affluent sections of the U.S. population, making its sample unrepresentative.

The *Gallup Poll* (one division of the Gallup Organization) conducts opinion polls in over 140 countries on an enormous range of political, economic and social issues. Its low point was probably in 1948, when it incorrectly forecast that Thomas Dewey would beat Harry S. Truman by a big margin in the U.S. presidential election. George Gallup believed the inaccuracy stemmed from ending his survey more than three weeks before the election.



George Gallup (1901–1984)

The aims of survey planning are to minimise both *costs* and *errors* (both sampling errors and non-sampling errors). These requirements are in conflict. Sampling error is reduced by choosing a larger sample, but costs are increased. We shall now briefly describe two further important tools of the survey planner's trade: first, a method of reducing sampling error (*stratified sampling*) and, second, a method of reducing costs (*cluster sampling*).



Despite producing results with no margin of error, the Eden poll is now defunct.

A method of sampling that reduces sampling error is often called **efficient**. This does *not* mean that it is cheap – such methods often cost more.

4.2 Stratified sampling

To reduce sampling error we have to reduce the potential variation between the different possible samples that we can choose. In other words, we want to make it more likely that the sample we choose is representative.

In Section 2 we assessed the representativeness of samples chosen from a listing of staff in the Sampling Department by analysing them with respect to gender and occupation. This was done by dividing the members of the sample into eight categories: these categories were the four occupational groups, each split into two genders. Having divided the sample into these eight categories, we then saw how the proportion of the sample in each category compared with the corresponding proportion for the whole population.

There were two reasons for using these particular eight categories for this analysis.

1. We knew the proportion of the whole population in each of these categories. (We could not base categories on salary levels, for instance, because they were not recorded on the list of the population that we had.)
2. It appeared likely that these categories were related to the subject of the investigation. To be more precise, it appeared that the data we collected from an individual on their economic well-being would depend on that individual's occupation and gender. It would not be possible to tell for certain if a sample was representative in terms of economic well-being without knowing the economic well-being of all the individuals in the population; and if we knew that, there would be no need to carry out the sample survey. But because economic well-being is thought to be related to occupation and gender, a sample that is representative in terms of occupation and gender is likely to be representative in terms of economic well-being too.

Categorising the population in this way is known as **stratification**: the eight categories are the **strata**. (A single category is a **stratum**.)

It is quite straightforward to ensure that any sample you might choose from the department is representative with respect to these eight strata. Instead of selecting members of the sample at random from the whole population, you would list the members of each stratum separately and then from each stratum select a number of individuals by simple or systematic random sampling. The selected individuals from a stratum form a **subsample**. You would then combine these subsamples (one subsample from each stratum) to get a sample from the whole population. This sample is then bound to be representative with respect to the strata, and is thus likely to be representative with respect to the subject of the investigation. Ideally, all the individuals in each stratum would be very similar to each other, so that whoever was selected from a stratum would be representative of that stratum. Then there would be comparatively little sampling error. A sample chosen in this way is a **stratified sample**.

This description of stratified sampling ignores one important point: how many individuals should be selected from each stratum, i.e. what should be the sizes of the subsamples? For example, suppose you want to deduce information about the average income of the members of the department (listed in Table 1) from data about the incomes of a sample. With a very small sample, there would not be much possibility of choice. With a sample of total size eight, you would have to choose a subsample of size one from each of the strata as it is an essential criterion of stratified sampling that in the sample there should be at least one member from each stratum.

However, if you are prepared to select a slightly larger sample, the ideas from Section 2 suggest that you should select approximately the same proportion of individuals from each stratum. If you wanted a sample of size 20 from the 86 members of the department, then you would select about the same proportion, $20/86$, of the people in each stratum.

For example, there are 17 people in the stratum of female secretarial staff; you might select about $20/86 \times 17$, which is about four, people from this category, and you might select ten or eleven men from the male professional category. You would still have to select the single male administrator and the male secretary.

Stratum subsample size

If approximately the same proportion of individuals are to be selected from each stratum, then

$$\text{stratum subsample size} \simeq \frac{\text{sample size} \times \text{stratum size}}{\text{total population size}}.$$

(Any subsample size less than one would be set equal to one.)

As described in Section 2, if you began by listing the population in order of strata (all the female professionals, then all the male professionals, followed by all the female administrators, the male administrator, and so on) and then chose a systematic random sample from the whole list, then the subsample sizes within each stratum would automatically come out to be approximately proportional to the stratum sizes.

However, when a little more is known about the population, it is sometimes better not to select a stratified sample in which the subsample sizes are proportional to the stratum sizes. For example, if you had the extra information that the incomes of the male professionals have a much larger spread than those of the female secretarial staff, then it would be better *not* to select the same proportion of each of these strata. This is because you need to obtain more information about the stratum with the larger spread in order to get the same amount of accuracy in your results. You should therefore choose a larger subsample from such a stratum, i.e. you should choose a relatively larger proportion of *male professionals* and a relatively smaller proportion of *female secretarial staff*.

This procedure makes more sense when we are thinking about sampling a large population, like electors in the UK, rather than a department with 86 people. With a large population, there would be thousands of people in each stratum, and it is easy to consider drawing subsamples whose size is proportional to the stratum size, or perhaps varying the proportions to take account of other available information.

In practice, most surveys that use subsample sizes which are not proportional to stratum sizes have a different reason for doing so. Suppose you were planning a survey of the adult population of England and Wales to investigate their subjective feelings on how well off they are. You would probably want to use stratified sampling, and you might well choose to stratify according to region of residence. You might work out that a total sample size of, say, 2000 would allow you to estimate sufficiently accurately what you want to know about the population of England and Wales as a whole. However, you might be particularly interested in comparing the results for Greater London with those for the rest of the country. Roughly

one-seventh of the population of England and Wales lives in Greater London, so if subsample sizes were chosen in proportion to stratum sizes, the Greater London subsample would consist of under 300 individuals. Such a sample size would probably not allow you to estimate sufficiently accurately what you want to know about the population of Greater London. You might therefore decide to increase the sample size for the Greater London subsample. In general, subsample sizes are often chosen so that appropriately accurate information is available on strata of particular interest, as well as for the population as a whole.

Stratification

Stratification is the categorisation of the population into strata that are:

- exhaustive: every member of the population must belong to a stratum
- mutually exclusive: no member of the population can belong to more than one stratum
- relevant to the subject under investigation: within each stratum, individuals should as far as possible be similar with respect to this subject
- known for all population members before the sample is chosen: otherwise a list of the individuals in a stratum from which to choose the subsample would not be available.

A stratified sample might then be chosen by selecting approximately the same proportion of individuals from each stratum. Such a stratified sample will be representative of the population with respect to the sizes of these strata. However, a stratified sample need not be chosen in this way, and often further knowledge about the population or the purpose of sampling will suggest better methods of selecting individuals from the strata.

These methods of stratified sampling ensure that the patterns in a stratified sample are less likely to be different from those in the population than are the patterns in a simple random sample of the same size. Therefore, the use of a stratified sample leads to more reliable results than the use of a simple random sample of the same size; in other words, the sampling error is reduced.

Example 4 *Survey of consumer prices*

You may remember from Unit 2 (Section 5) that the calculation of the RPI uses a monthly survey of retail prices carried out by a market research company on behalf of the UK Office for National Statistics. In this survey, prices are collected from a sample of shops situated in approximately 150 locations across the UK. This sample of shops is stratified: each shop is put into one stratum according to which of the 12 regions of the country it is in and which of the three types of retail outlet it is.

Activity 11 *Who bought the seed?*

Suppose that you work for a mail order seed company and you wish to carry out a sample survey of the population of UK customers who bought seed of a new variety of pea to find out their opinion of it. You have computerised records of the names and addresses of all these customers, and of the amount of seed of this variety that each of them bought. How would you go about dividing this population into strata?

Stratified sampling has one disadvantage which is normally relatively minor: it can increase costs. This is because to use this method it is necessary to spend time discovering information about the population and then carefully distinguishing the strata and deciding the subsample sizes. We shall now look at a method which, in contrast, can produce dramatic savings in costs in certain types of survey.

4.3 Cluster sampling

Many surveys involve interviewers contacting individual members of the chosen sample in their homes or work places. A survey of this kind can be enormously expensive, particularly if it covers a wide geographical area such as the whole of the UK, because the interviewers' travel time and transport costs are both considerable. It is obviously in the interests of economy to arrange, if possible, for the individual members of the sample to be not too widely dispersed geographically.

Here, then, is a brief description of **cluster sampling**: a method that cuts the costs of such surveys by restricting the sample to a limited number of geographical areas.

One major survey that involves such personal interviews is the Living Costs and Food Survey.

Choosing a cluster sample

Cluster sampling works as follows:

1. Find suitable geographical areas.
2. Choose, preferably using random methods, a limited number of these geographical areas.
3. For each of these chosen geographical areas, choose a subsample from those members of the population in that area.
4. Combine these subsamples (one from each chosen area) to get a sample.

The population in each geographical area is a **cluster**, and such a sample is a **cluster sample**. Clusters may also consist of entities other than geographical areas.

For this method of cluster sampling to produce representative samples, it is essential that the populations in the chosen clusters are, between them, representative of the whole target population.

In 1947, Hollywood made a film (*Magic Town*, starring James Stewart) about a small town in the Midwest of the USA which was a microcosm of American Society. This single town of about 2000 inhabitants was found to represent the whole country in its social, economic and political characteristics. Any such town, in any country, would be ideal for official surveys, for market research and for public opinion polls because all such surveys could confine their attention to a sample from this one town, i.e. they could choose just one cluster. A few hours' work interviewing a random sample of individuals from this town would produce representative results about the whole population of the country, saving enormous amounts of time and money. Such towns, however, exist only in a Hollywood producer's imagination. The real world is no Hollywood! Towns within a country differ quite a lot in their characteristics, depending, for example, upon where they are, the age of their populations and the major local employers.

For this reason, it is never sensible to confine a cluster sample to a single cluster. The usual practice is to choose several clusters using random sampling; then a subsample is selected from each chosen cluster, again normally by simple random sampling.

There are various forms of cluster sampling. One form is described below.

One form of cluster sampling

1. Specify the number of clusters to use in the survey and the proportion that is to be surveyed from each of the selected clusters.
2. Choose which clusters to use at random, with each cluster having the same probability of being included in the survey.
3. Draw a simple random sample from each of these clusters. The clusters may differ in their sizes, and the sizes of the subsamples drawn from them should vary correspondingly: subsample approximately the same pre-specified proportion of each cluster.

A desirable property held by this form of cluster sampling is that every individual in the target population has approximately the same probability of being included in the survey. (Small differences between the probabilities will usually be inevitable because the sample sizes must be whole numbers.) A drawback, though, is that the total sample size will partly depend on which clusters are chosen – if large clusters are chosen by chance in step (b), then the total sample size will be larger than when step (b) yields small clusters. There are forms of cluster sampling that avoid this drawback, but we will not consider them in M140.

Although cluster sampling saves money, it also has a disadvantage: other things being equal, cluster sampling will almost always lead to greater sampling errors than would arise in a simple random sample of the same size. The reason for this is that individuals within clusters tend to be less variable than individuals in the target population as a whole. Two people living in the same town are likely to be more similar than two people living in different towns. By restricting the sample to the chosen clusters, it is thus likely to be less representative.

However, suppose a survey is being planned within a fixed budget. Very often the cost savings achieved by using clustering allow the sample size to be increased to such an extent that the results from the cluster sample are considerably more reliable than the results would be from the very much smaller unclustered sample that could be afforded.

Do not forget that this argument applies only to surveys using interviewers who have to travel. It would not apply, for example, in a survey carried out by post. For many such postal surveys, there is no reason for using clustering on a geographical basis. However, there is another good reason for using cluster sampling in some situations. To draw a simple random sample, a complete list of the target population is required. For some populations, it would be a major undertaking to produce such a list. No complete single listing of all UK schoolchildren exists, for instance, and it would not be feasible to produce one. It would be much more feasible, for a survey of this population, to obtain a list of all schools, to choose a

There are circumstances in which cluster sampling is likely to produce a sample that is more representative than a simple random sample of the same size, but in practice these circumstances hardly ever arise.

limited number of schools as clusters, make a list of the pupils in each of the selected schools, and draw samples from these lists.

Although cluster sampling usually makes use of geographical areas, there are other ways of dividing a population into clusters. For example, suppose a chocolate manufacturer wanted to sample his chocolates at the end of production, in order to test for quality. It would be economical to select boxes of chocolates at random and then to select several (or perhaps, all) of the chocolates from the selected boxes for testing. This would avoid wasting too many boxes. Here, each box of chocolates is a cluster.



4.4 Stratified and cluster sampling

Let us now summarise the main points from the last three subsections and compare these two methods.

Stratified sampling

- Each stratum focuses on one section of the population, such as those of a specified gender in a particular age group.
- Every member of the population must be in one and only one stratum.
- A stratified sample includes members of every stratum.
- Stratified sampling decreases sampling error compared to a simple random sample of the same size (i.e. it is more efficient) but slightly increases costs.

Cluster sampling

- Each cluster should be, as far as possible, a representative cross-section of the whole population.
- Every member of the population must be in one and only one cluster.
- A cluster sample excludes all the members of some (usually most) of the clusters.
- Cluster sampling often decreases costs but usually increases sampling error compared to a simple random sample of the same size (i.e. it is less efficient).

Many well-planned surveys use both strata and clusters. An example of such a survey is the Living Costs and Food Survey, introduced in Unit 2. There are also elements of both in quota sampling, as you will see in the next subsection.

4.5 Quota sampling

Quota sampling is a procedure that is used frequently for market research surveys and opinion polls. Firstly the sample size is determined (usually by consideration of costs), and then each interviewer is allocated a quota of interviews to achieve. The interviewers are then sent out to contact suitable respondents at selected sites in selected towns (Figure 11).



Figure 11 Data collection

These sites might be supermarkets, railway stations, high streets, etc. Thus the quota sample is a cluster sample. The sample is stratified by requiring interviewers to interview a fixed number of people in specific groups such as age, gender and occupation groups.

A quota sample is not a random sample: the selection of individuals is haphazard rather than random.

Quota sampling is economical because it produces quick results. These results are, however, often of dubious reliability because the method can introduce error. Market researchers are fond of quoting the apocryphal story of the interviewer who quickly achieved his full quota of interviews from people queuing for a train at Liverpool Street Station in London. The survey was about gambling and all those interviewed were waiting for a special train to take them to the Newmarket horse races!

Another disadvantage of quota sampling is that it is usually difficult to give a numerical estimate for how unrepresentative the results are likely to be. It is possible to give such estimates for random sampling methods, using the ideas of probability that you will meet in Unit 6.

4.6 Sampling from the electoral register

Most of the methods of choosing a sample described in this unit require a list of all the individuals in the target population. This list is sometimes called the **sampling frame**. One sampling frame that has commonly been used in the UK for surveys of individual adults and of households is the **electoral register** (such as that shown in Figure 12). This lists all electors and it is possible to buy an edited version. The full register contains almost all adults who are eligible to vote, as the registration of eligible voters is compulsory in the UK. However it does not contain many non-EU citizens or any people aged under 17. (People can be registered to vote from age 17, though their registration is not activated until they reach their 18th birthday.) Also, the edited register does not include anybody who has chosen not to be included in the edited version. Another drawback of the electoral register is that it is out-of-date even when it is first published, because compiling a relatively complete list of a large human population is time-consuming.

Voter's Surname	Voter's Christian Name	Description and Situation of Rateable Property	Number of Votes to which Voter is entitled	Electoral District in which Property is situated	Division of Electoral District in which Property is situated
Atkins	John	Land	One	North Gipping Land	Middle
Allen	Thomas	Land	One	do.	do.
Anderson	John	House and land	One	do.	do.
Anarbyce	John	Land	Two	do.	do.
Bently	Joseph	Land	One	do.	do.
Barker	Robert	Land	One	do.	do.
Baylies	Charles	Land	One	do.	do.

Figure 12 Avon Roll 1866

We will use the electoral register for a part of Milton Keynes to illustrate some of the survey methods that have been discussed. We will suppose the target population is the adult residents of five streets (Jersey Close, Kerrera Close, Lytham Gardens, Melton Gardens and Norfolk Place) and that the purpose of the survey is to learn about their bus usage. People participating in the survey will be asked:

Did you use a bus service in Milton Keynes in the last week?

Table 11 lists the adults living in the target streets, based on the electoral register. It also records whether or not they had used a bus service in Milton Keynes during the week – though this information would only be known for those people questioned in the survey.

Table 11 Bus usage in a part of Milton Keynes

Registration number	Name	Street number	Bus user?
Jersey Close			
977	Denton, George	1	Y
978	Wells, Joan F	2	Y
979	Hanrahan, Brian K	3	N
980	No Elector	4	
981	Jones, Ian	5	Y
982	Jones, Linda	5	Y
983	Abbott, David	6	N
984	Abbott, Mary R	6	Y
985	Donegan, Andrew B	7	N
986	Donegan, Margaret H	7	N
987	Turner, Thomas J F	8	Y
988	Turner, Florence P	8	Y
989	West, Michael J	9	Y
990	West, Jean P	9	N
991	Nelson, Sheila A	9	N
992	Mason, Arthur B	10	N
993	Mason, Joan M	10	Y
994	Wilson, Annabel N	11	N
995	Wilson, Lillian	11	Y
996	Chapman, Reginald R	12	Y
997	Chapman, Iris	12	Y
998	Watson, Richard T	12	N
999	No Elector	13	
1000	Mercer, Gladys C	14	Y
Kerrera Close			
1001	Groves, Jacqueline F	1	Y
1002	Drinkwater, James G	1	Y
1003	Tong, Michael	2	N
1004	Burton, Christopher N	3	Y
1005	Hexton, Amara	4	N
1006	Hexton, John	4	Y
1007	Smith, Alan C	5	Y
1008	Dixon, Mary C	6	Y
1009	Daly, Sean	6	Y
1010	Ho, Audrey	7	N
1011	Tongwell, Kim	8	N
1012	Clark, Michael E	9	N
1013	Clark, Jennifer	9	Y
1014	Christon, John E	10	Y
1015	Christon, Clare M	10	Y
1016	Dunn, Garry A	11	Y
1017	Dunn, Mary E	11	Y
1018	Edwards, Kathleen	12	N
1019	Edwards, Vince L	12	Y
1020	Price, Eleanor T	12	N
1021	Goulding, Matthew M	13	Y
1022	Goulding, Janet	13	Y
1023	Turner, Lee	14	Y
1024	Bailey, Ivy W	15	Y
1025	McCann, Raymond D	16	Y
1026	McCann, Victoria K	16	Y
1027	Wyatt, Edith	17	N

Registration number	Name	Street number	Bus user?
Lytham Gardens			
1028	Kerr, John M B	1	<i>N</i>
1029	Kerr, Susan	1	<i>N</i>
1030	Kerr, Lynn	1	<i>Y</i>
1031	Kerr, David	1	<i>Y</i>
1032	Kohler, Martina	2	<i>Y</i>
1033	Kohler, Nicholas	2	<i>N</i>
1034	Clements, Neil S	3	<i>N</i>
1035	Clements, Marie A	3	<i>N</i>
1036	Clements, Ian P	3	<i>N</i>
1037	Patel, Suresh	4	<i>Y</i>
1038	Knight, Patricia H	5	<i>N</i>
1039	Bolton, Samuel T	6	<i>N</i>
Melton Gardens			
1040	Clarke, David P	1	<i>N</i>
1041	Clarke, Annette M L	1	<i>N</i>
1042	Barnard, Ruby	2	<i>N</i>
1043	No Elector	3	
1044	French, Richard E	4	<i>N</i>
1045	Coe, Alanah	4	<i>Y</i>
1046	Smith, Angela	5	<i>Y</i>
1047	Ferguson, Brian	6	<i>N</i>
1048	Ferguson, Sally	6	<i>N</i>
1049	Ferguson Michael	6	<i>Y</i>
1050	Shah, Jaya	7	<i>Y</i>
1051	O'Neill, Thomas	8	<i>Y</i>
1052	O'Neill, Mary S	8	<i>Y</i>
1053	Hedley, Robert M	8	<i>N</i>
1054	Scott, Ian R	9	<i>Y</i>
1055	Scott, Dorothy G	9	<i>N</i>
1056	McGregor, David E	10	<i>N</i>
1057	McGregor, Aileen J	10	<i>N</i>
1058	Paine, Darrell R	11	<i>N</i>
1059	Paine, Lynne C	11	<i>N</i>
Norfolk Place			
1060	Fisk, Catherine A	1	<i>N</i>
1061	Hatley, Brian J	2	<i>N</i>
1062	Brooke, Denise	2	<i>Y</i>
1063	Lang, Deborah M	3	<i>Y</i>
1064	Flynn, Horace I	4	<i>N</i>
1065	Flynn, Ann C	4	<i>Y</i>
1066	Shah, Dipak	5	<i>Y</i>
1067	Shah, Mala	5	<i>N</i>
1068	McTaggart, William E	6	<i>Y</i>
1069	McTaggart, Christine V	6	<i>N</i>
1070	McTaggart, James J	6	<i>N</i>
1071	Hall, Stephen D	7	<i>N</i>
1072	Godman, Janet K	7	<i>N</i>
1073	Weston, Zoe	8	<i>N</i>
1074	Uttley, Muriel O	9	<i>Y</i>

To sample from the electoral list in Table 11, we use random numbers and relate these to the registration numbers. The registration numbers run from 977 to 1074, so, with random number tables, it is efficient to use pairs of random digits:

- 79 would mean ‘Registration number 979’
- 73 would mean ‘Registration number 1073’
- 00 would mean ‘Registration number 1000’.

We would ignore 75 and 76, and also the pairs corresponding to ‘No Elector’.

Example 5 *Simple random sample of 12 electors*

Suppose a simple random sample of twelve electors is required. If we use the random number table in the appendix and start at the beginning of row 49, then the selected random numbers are:

96, 00, 26, 82, 60, 22, 02, 60, 69, 99, 09, 67, 01, 12, 01, . . .

Equating these to the corresponding electoral registration numbers determines our sample. (The second 60 will be ignored, because we want no repeats, and 99 will be ignored because 999 is a ‘No Elector’.) The electors in the sample and their characteristics are given in Table 12.

Table 12 Bus-usage in a simple random sample

Registration number	Name	Address	Bus user?
996	Chapman, Reginald R	12 Jersey Close	Y
1000	Mercer, Gladys C	14 Jersey Close	Y
1026	McCann, Victoria K	16 Kerrera Close	Y
982	Jones, Linda	5 Jersey Close	Y
1060	Fisk, Catherine A	1 Norfolk Place	N
1022	Goulding, Janet	13 Kerrera Close	Y
1002	Drinkwater, James G	1 Kerrera Close	Y
1069	McTaggart, Christine V	6 Norfolk Place	N
1009	Daly, Sean	6 Kerrera Close	Y
1067	Shah, Mala	5 Norfolk Place	N
1001	Groves, Jacqueline F	1 Kerrera Close	Y
1012	Clark, Michael E	9 Kerrera Close	N

Eight individuals in this sample of 12 people are bus users, so the sample estimate of the percentage of bus users in the population is

$$\frac{8}{12} \times 100\% \simeq 66.7\%.$$

In the target population of 95 electors, there are actually 49 people who used the bus in the previous week, so the true percentage of bus users is $49/95 \simeq 51.6\%$.

Activity 12 *Systematic random sample*

Suppose a systematic random sample of about one-eighth of the targeted electors is required. Select such a sample, taking the first digit in the range 1 to 8 from row **6** as a random start. List the names of the electors in the sample and whether they are bus users. Based on this sample, what is the estimated percentage of bus users in the target population?

Some sampling schemes divide the population into categories that are sampled separately. (Some categories might not be sampled, as in cluster sampling, for example, where only selected clusters are sampled.) Having chosen the categories to sample, each category is taken in turn and a simple random sample drawn from it.

Example 6 *Stratified sample of 12 electors from two strata*

Suppose the five streets that give our target population can be sensibly divided into two strata: Jersey Close and Kerrera Close were, at the time of the survey, both newly built and form one stratum, while Lytham Gardens, Melton Gardens and Norfolk Place were all built about twenty years earlier and form a second stratum. The strata are of similar size (49 electors in one stratum and 46 in the other), so we will sample the same number of people from each stratum, i.e. six from each.

We will start in row **16** of the random number table.

16	47 14 97	61 57 30	93 88 12	88 58 15	75 17 45
17	98 75 58	14 05 05	16 72 57	34 20 46	91 04 44
18	64 71 77	50 51 00	61 02 60	51 13 61	34 33 73

The electoral registration numbers for the first stratum range from 977 to 1027, so we look through the random numbers picking out those between 77 and 99, and those between 00 and 27, but ignore duplicates and those corresponding to 'No Elector'. This gives 14, 97, 93, 88, 12 and 15. For the second stratum, we start reading random numbers from where the previous sample ended, picking out those between 28 and 74: 45, 58, 72, 57, 34 and 46. The electors corresponding to these numbers, together with their characteristics, are listed by stratum in Table 13.

Table 13 Bus usage in a stratified sample

Registration number	Name	Address	Bus user?
Jersey Close and Kerrera Close			
1014	Christon, John E	10 Kerrera Close	Y
997	Chapman, Iris	12 Jersey Close	Y
993	Mason, Joan M	10 Jersey Close	Y
988	Turner, Florence P	8 Jersey Close	Y
1012	Clark, Michael E	9 Kerrera Close	N
1015	Christon, Clare M	10 Kerrera Close	Y
Lytham Gardens, Melton Gardens and Norfolk Place			
1045	Coe, Alanah	4 Melton Gardens	Y
1058	Paine, Darrell R	11 Melton Gardens	N
1072	Godman, Janet K	7 Norfolk Place	N
1057	McGregor, Aileen J	10 Melton Gardens	N
1034	Clements, Neil S	3 Lytham Gardens	N
1046	Smith, Angela	5 Melton Gardens	Y

Seven individuals in this sample of 12 people are bus users, so this sample estimates the percentage of bus users in the population as

$$\frac{7}{12} \times 100\% \simeq 58.3\%.$$



Example 6 is the subject of Screencast 4 for Unit 4 (see the M140 website).



Activity 13 Cluster sampling with subsamples of one-third

Suppose that the streets in the population listed in Table 11 were widely separated geographically, and that therefore you wanted to use cluster sampling for your survey, restricting your sample to just two of the streets and sampling approximately one-third of the individuals in each cluster. Obtain the sample using the following procedure:

- Number the streets from 1 to 5 in the order in which they are listed. Using *single* random digits, and starting at the beginning of row **26** of the random number table in the appendix, select the two streets to be sampled. These streets are to be sampled in the order in which they are selected.
- Determine the sizes of the samples to take from each cluster (street) by dividing each cluster size by 3 and rounding the results *up* to whole numbers.
- To select individuals for the subsample from the first selected street, use pairs of digits starting at row **82** of the random number table. No person may be selected more than once. To select individuals from the second subsample, continue from the point reached in the random number table after selecting the first subsample, and apply the same procedure again.

List the people chosen in the subsamples and estimate the proportion of bus users in the target population.

Activity 13 is the subject of Screencast 5 for Unit 4 (see the M140 website).



4.7 Some more considerations

Even if you ever thought that sampling would be child's play, you should now be able to appreciate that it is a good deal more difficult than pulling rabbits out of hats, and in addition, that it can involve a lot of hard slog. Here are a few more of the problems that abound in this work.

- **Defining the target population.** Sometimes this is not straightforward. For example, in an opinion poll designed to predict the result of an election, the target population is all those people who will actually vote on polling day, but who these people are cannot be known beforehand.
- **Listing the target population.** Most of the methods of choosing a sample described in this unit require a sampling frame. (An advantage of cluster sampling is that it does not require a full sampling frame.) It is often difficult to obtain an accurate list, as you saw in the description of sampling from the electoral register.
- **Non-contact and non-response.** Often it is impossible to contact everyone in the sample, and some of the individuals contacted may not be able or willing to provide the required information.
- **Questionnaire design.** This could well be the subject of a whole unit. Devising questions that will discover the required information is not easy. Also, for example, the way in which the questions are asked by the interviewer may well affect the answer.
- **Clerical errors.** No matter how carefully the work is done there are certain to be errors in recording and transcribing the data. Many of these will, however, be discovered if the data are analysed sensibly.

In this section, you have read about the principles involved in cluster sampling, stratified sampling and quota sampling. You now know about some of the problems in sampling, and in particular some problems of sampling from the electoral register.

Exercises on Section 4

These exercises consider how sampling might be used to investigate households whose expenditure may not fit typical patterns used by the Retail Prices Index (RPI).

Exercise 6 *Cluster sampling?*

Households which own their home outright, and therefore do not make either mortgage or rent payments, might well have a considerably different expenditure pattern to other households, and the RPI may therefore not be an accurate indicator of inflation as they experience it, particularly as the Housing sub-group has the highest weight in the RPI. Suppose you are required to select a national sample of such households so that their expenditure can be analysed separately.

- (a) State, with a reason, whether cluster sampling would be a valid and appropriate method to use for the initial stage of selecting such a sample.
- (b) Explain which method of sampling you would use to select the individual households in your final sample, justifying your choice of method.

Exercise 7 *Sampling methods and sampling frames*

The Motoring Expenditure sub-group has the second-highest weight in the RPI. In some rural areas, households which do not own a motor vehicle, and are therefore dependent on public transport, may have a different expenditure pattern to the majority of households that do own a vehicle. The RPI may therefore not be an accurate indicator of inflation as experienced by rural households without a vehicle. Suppose you are required to select a national sample of such households so that their expenditure can be analysed separately.

- (a) A pilot survey is to be carried out in one area. What official records might you want to access to obtain a suitable sampling frame from which a sample of such households could be obtained?
- (b) State which sampling method you would use to select the sample from the sampling frame, justifying your choice.



Exercise 8 *Stratified sampling*

Suppose the electorate given in Table 11 divides into three strata: Jersey Close, Kerrera Close and the other three roads. A random sample of size 12 is to be drawn from this population using stratified random sampling.

- (a) Select the subsample sizes so that they are approximately proportional to the stratum sizes, ensuring that the total sample size is 12.
 - (b) Select the sample, using simple random sampling from each stratum in turn. Start at the beginning of row **52** of the random number table in the appendix. Write down the names of the electors you select and whether or not they are bus users.
 - (c) Calculate the percentage of electors sampled who are bus users and comment briefly on how well your sample represents the target population (of all adults living in this part of Milton Keynes) in terms of using the bus service.
-

5 Computer work: sampling



In Section 3, you looked at sampling from a target population and learned about sampling distributions. In this section, you will explore the sampling distribution for samples of size 3 taken from a particular target population, followed by looking at sampling distributions for samples of different sizes. You will then learn how to use Minitab to produce simple random samples.

You should now turn to the Computer Book and work through Subsection 4.1, if you have not already done so, followed by the rest of Chapter 4.

Summary

This unit has focused on statistical issues surrounding one method of data collection – surveys. In a survey, information is collected about a sample of individuals and used to draw conclusions about the population as a whole. Different methods are used to select samples, the best method depending on the survey and the target population.

- In simple random sampling, every possible sample of a given size has an equal chance of being selected. This is usually done by selecting individuals at random from the population. In systematic random sampling, individuals are chosen by working systematically down a list, with only the starting point chosen at random.
- Stratified sampling and cluster sampling assume that the population can be split into groups. In stratified sampling, individuals from every group are selected, ensuring that every group is represented in the sample. In cluster sampling, individuals in the sample only come from selected groups, ensuring that sampling process is more cost-efficient.
- In quota sampling, individuals are not selected at random, though they are chosen so that different groups in the population are represented fairly.

You have also learned in this unit about the sampling distribution of the median. That is, how the sample median varies according to which particular sample happened to be selected. You have seen that the sample median is not necessarily equal to the population median, even when there are just five categories to choose from. Indeed when the sample size is very small, it might be more likely to be different to the population median. However as the sample size increases, it becomes more likely that the sample median is the same as the population median.

Learning outcomes

After working through this unit, you should be able to:

- explain in general terms why a well-chosen sample is an economic and accurate method of collecting data about a population
- choose a simple random sample using random numbers and a labelled list of the target population
- choose a systematic random sample using random numbers and a labelled list of the target population
- describe the differences between, and outline the relative strengths and weaknesses of, simple and systematic random sampling
- give an example of the type of pattern that can be seen in the collection of all possible samples of a given size
- interpret descriptions and pictures of distributions which are expressed in proportions
- describe the principles involved in cluster sampling and stratified sampling
- describe quota sampling in general terms
- choose a random sample for a stratified survey using random numbers and a labelled list of the target population
- choose a random sample for cluster sampling using random numbers and a labelled list of the target population
- describe some of the problems in sampling, and in particular some problems of sampling from the electoral register.

Appendix: random number table

This table contains 3000 random digits (i.e. throws of a ten-sided die labelled 0, 1, ..., 9).

1	98 06 77	46 16 63	99 80 81	82 15 48	96 12 56	51	64 60 21	12 41 60	04 63 93	45 25 52	75 50 35
2	41 25 66	21 51 66	11 34 33	18 36 41	33 18 70	52	64 76 41	17 07 54	01 29 86	41 93 16	55 54 40
3	68 58 71	24 92 06	94 84 48	92 96 32	29 00 60	53	93 46 82	67 64 48	91 74 85	94 40 51	30 93 08
4	78 32 89	76 61 03	01 20 94	36 39 87	52 27 23	54	82 64 44	58 45 94	30 39 86	19 64 84	35 30 19
5	50 76 11	36 13 84	32 93 72	29 04 41	25 43 89	55	61 46 40	89 21 47	20 85 91	90 56 67	40 31 46
6	71 58 45	43 72 69	18 67 32	57 29 57	02 58 68	56	92 80 33	89 23 96	24 33 16	80 45 20	35 36 00
7	82 14 64	47 40 74	53 03 75	40 28 63	53 36 90	57	45 65 20	02 56 40	21 35 17	71 33 07	36 71 90
8	41 01 53	67 41 78	84 29 26	34 42 19	82 31 79	58	40 99 02	66 37 59	24 79 35	21 61 29	96 50 01
9	30 63 22	27 28 69	36 23 99	52 29 03	87 28 54	59	50 31 47	84 44 30	70 33 12	63 54 86	63 08 62
10	73 09 03	54 20 02	55 49 48	46 75 42	62 63 42	60	87 38 68	91 50 98	65 95 29	54 20 89	25 59 33
11	66 41 48	46 17 24	82 51 86	86 53 66	95 57 95	61	20 85 20	04 05 21	53 58 65	40 60 53	91 07 32
12	70 10 21	02 71 89	14 80 64	32 58 17	35 65 55	62	03 16 83	56 90 30	18 77 83	18 91 26	70 50 99
13	59 55 94	44 77 90	01 99 79	48 28 61	93 87 17	63	57 12 84	22 89 61	19 55 62	96 01 44	28 96 21
14	75 81 42	45 69 28	23 90 46	24 32 97	64 41 70	64	11 06 38	37 58 65	66 54 73	80 38 57	44 99 81
15	72 22 05	84 39 89	57 73 84	86 57 76	79 08 65	65	44 68 39	54 96 66	56 83 21	22 34 00	28 65 20
16	47 14 97	61 57 30	93 88 12	88 58 15	75 17 45	66	73 19 05	41 32 92	36 98 10	94 60 47	32 10 82
17	98 75 58	14 05 05	16 72 57	34 20 46	91 04 44	67	39 56 14	02 45 65	16 86 78	90 46 39	58 62 66
18	64 71 77	50 51 00	61 02 60	51 13 61	34 33 73	68	32 53 16	30 76 36	80 52 65	02 10 07	81 40 80
19	43 12 15	66 40 56	39 77 75	32 80 30	22 90 95	69	98 43 67	05 82 06	19 24 86	24 30 44	06 15 54
20	59 70 46	36 67 19	12 59 39	42 35 24	69 86 14	70	53 08 00	94 46 80	60 94 01	83 94 45	42 43 55
21	25 84 20	27 35 05	54 21 39	04 77 69	78 76 99	71	28 21 05	43 60 40	73 70 75	33 10 74	91 83 95
22	40 52 36	07 18 99	79 27 36	30 97 14	72 64 82	72	89 79 63	50 98 53	56 42 12	76 48 56	34 46 82
23	48 38 90	79 26 63	50 41 87	76 31 13	81 55 34	73	61 48 17	25 59 95	19 14 31	68 94 23	83 40 83
24	61 91 66	85 68 10	40 47 44	71 56 81	00 34 07	74	41 98 20	72 70 69	39 46 17	37 70 37	81 75 23
25	45 40 26	25 37 27	02 15 26	27 51 87	18 91 30	75	51 08 35	35 16 20	92 94 25	05 04 01	65 33 82
26	32 57 79	72 02 27	96 10 62	63 07 30	01 40 97	76	73 97 76	94 92 07	24 89 41	98 35 91	96 52 82
27	69 23 49	01 02 17	28 23 72	71 46 39	24 46 39	77	43 74 49	01 59 38	60 29 94	61 02 11	61 86 36
28	63 80 25	47 36 69	73 39 21	23 93 10	09 50 45	78	94 94 39	87 49 44	54 02 52	56 28 49	34 49 25
29	31 30 49	19 65 12	33 87 76	64 22 62	66 61 88	79	52 10 65	11 34 68	68 65 58	90 17 33	98 36 82
30	68 42 66	14 60 63	24 06 92	94 21 52	71 37 19	80	54 42 73	62 51 54	80 63 36	65 12 44	52 16 12
31	52 77 76	33 55 75	78 03 11	18 04 23	12 72 46	81	73 27 51	94 71 14	37 55 00	05 32 36	59 89 86
32	19 05 93	62 41 96	47 15 34	80 17 23	06 44 75	82	77 69 59	62 33 99	26 67 95	72 77 16	02 28 96
33	15 23 16	85 63 28	62 03 72	11 74 17	35 37 09	83	08 19 98	26 68 06	02 05 57	21 73 55	35 07 79
34	32 84 18	60 89 57	09 25 31	82 79 92	10 08 71	84	50 83 92	60 44 28	52 83 25	39 83 60	92 71 10
35	59 10 86	85 92 14	14 17 38	59 35 24	12 53 88	85	16 89 30	82 48 70	63 82 71	48 72 82	77 37 56
36	18 56 17	74 42 45	19 35 75	18 37 47	42 78 08	86	21 41 74	65 08 73	82 94 72	22 67 92	34 74 33
37	28 87 01	51 67 42	00 77 30	16 31 06	67 42 75	87	99 08 47	77 43 94	17 07 76	57 93 68	61 15 97
38	83 25 37	02 91 92	05 16 09	07 35 84	59 15 44	88	20 02 69	70 87 44	57 23 35	99 94 16	63 40 99
39	12 09 73	08 61 72	89 23 91	85 76 99	29 55 48	89	93 95 15	81 21 75	71 39 23	31 06 43	87 44 21
40	64 74 95	68 36 68	69 99 56	33 78 08	84 31 87	90	10 91 65	40 88 43	50 57 83	50 82 34	12 78 80
41	77 46 18	83 52 40	05 76 20	95 40 64	73 67 44	91	91 72 35	36 80 19	49 49 37	17 40 98	02 53 59
42	06 69 75	42 75 68	99 14 90	83 26 03	15 00 71	92	23 82 82	20 56 34	76 49 27	40 78 29	99 07 22
43	75 53 11	01 11 11	78 56 62	03 87 34	18 12 42	93	21 02 08	25 07 15	36 45 19	21 30 48	30 76 99
44	09 30 87	33 32 37	96 79 07	33 75 21	74 06 47	94	80 22 31	36 25 82	63 91 94	56 59 42	34 18 65
45	02 30 44	66 34 64	38 75 01	40 22 87	76 19 01	95	46 92 44	39 46 22	03 99 15	60 45 34	06 77 86
46	14 45 74	30 52 97	77 13 20	66 87 54	89 05 30	96	54 81 74	93 71 51	14 28 22	66 21 53	31 66 09
47	82 45 49	85 02 33	58 84 03	74 63 52	15 47 04	97	80 17 11	33 37 07	00 77 89	31 86 72	35 67 41
48	44 33 94	98 75 51	62 00 17	59 00 42	09 39 66	98	05 47 12	99 05 06	18 52 83	53 36 90	84 63 41
49	96 00 26	82 60 22	02 60 69	99 09 67	01 12 01	99	78 99 91	58 03 59	93 60 31	40 23 58	35 51 66
50	20 67 56	12 77 16	78 04 36	38 95 35	71 26 49	100	20 33 54	25 07 06	55 95 53	14 64 58	01 07 63

Solutions to activities

Solution to Activity 1

The labels selected are

52 10 65 11 34 68 58 90 17 33 98 36.

The list cannot be obtained by simply taking the first 12 pairs along the row; the seventh pair is 68, which has already appeared in the sample, and the eighth pair is 65 which has also already appeared, so you should have ignored the seventh and eighth pairs.

Solution to Activity 2

The labels selected are

72 22 05 84 39 89 57 73 84 86 57 76 79 08 65
47 14 97 61 57 30 93 88 12 88 58 15 75 17 45.

This time there is no problem with repeated individuals in the sample.

Solution to Activity 3

The sample is shown in the following table:

Name	Label	Gender	Occupation
Hare, Dorothy	41	F	P
Dev, Mohen	25	M	P
Redman, Guy	66	M	P
Crofts, Mary	21	F	A
Lang, Chris	51	M	P
Bramley, Max	11	M	P
Graham, Bert	34	M	P
Gowan, Dai	33	M	P
Cluskie, Alex	18	M	P
Grant, Lynne	36	F	P
Rowan, George	70	M	P
Ricardo, Dan	68	M	P
Masterton, Dick	58	M	P
Sandford, Dave	71	M	P
Damper, Emma	24	F	S
Bates, Sheila	06	F	S
Woodhouse, Paul	84	M	M
James, Patricia	48	F	A
Franks, Abraham	32	M	P
Fallow, Jim	29	M	P

The sample of size 20 that you have just obtained is rather more representative of the population than was the previous sample of size 10. In this larger sample, 70% are men and 30% women, compared to 60% and 40% in the population. In addition, this sample fairly closely represents the occupational pattern in the population. It slightly over-represents the professional staff and under-represents secretarial staff. In a sample of size 20, you might expect about four secretarial staff; this sample has only

two. However, this larger sample should represent the population quite well for most practical purposes. (That is not to say, of course, that every simple random sample of size 20 would represent the population as well!)

Solution to Activity 4

Step 1 The first pair of digits from row **3** in the range 01 to 17 is 06, so this is the random start. (Notice that you must use pairs of digits; you cannot use the single digit 6 at the beginning of the line.)

Step 2 The labels in the sample are every 17th label:

06 23 40 57 74.

The sample is shown below.

Name	Label	Gender	Occupation
Bates, Sheila	06	F	S
Daley, Stuart	23	M	P
Hallow, Jean	40	F	A
McCraig, Frank	57	M	P
Stratford, Peter	74	M	P

For such a small sample, this is about as representative of the target population as you might hope. There are three men and two women, which is the same ratio as in the population. Also, there are three professionals, one member of the secretarial staff and one administrator; this is a fair representation of three of the categories. There are no manual workers in this particular sample.

Solution to Activity 5

Step 1 The first digit in row **29** is 3, so we start at label 03.

Step 2 The labels in the sample are every fourth label:

03 07 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83.

The sample is shown in the following table.

Name	Label	Gender	Occupation
Archer, Simon	03	M	M
Baxter, John	07	M	P
Bramley, Max	11	M	P
Chapman, Liz	15	F	M
Cramer, Will	19	M	P
Daley, Stuart	23	M	P
Eric, Steve	27	M	P
Foster, Sue	31	F	S
Graham, Bill	35	M	P
Greenway, Maggie	39	F	P
Hewitt, Ray	43	M	P
Iron, Donald	47	M	P
Lang, Chris	51	M	P
Lupton, David	55	M	P
Menton, Christine	59	F	S
Osterley, Rebecca	63	F	S
Redstar, Pamela	67	F	S
Sandford, Dave	71	M	P
Thompson, Anna	75	F	S
Turner, Richard	79	M	P
Winston, Chuck	83	M	P

There are 14 men in the sample of 21, which is 67% compared to 59% of the target population. There are also 14 professionals (67%) compared to 65% in the target population. 24% of the sample are secretarial staff, compared with 21% of the population. There are two manual workers but no administrators. On the whole, this sample provides quite a good representation of the target population. The lack of representativeness is not really any more than one might expect in a sample of this size.

Solution to Activity 6

Since the batch size is 1000, the median is halfway between the 500th and 501st values. Counting in 500 from the ‘Much worse off’ end of the population, responses 1 and 2 (‘Much worse off’ and ‘Somewhat worse off’) account for 400 values, so the 500th value is 3. Similarly the 501st value is also 3, so the median is 3.

Solution to Activity 7

If we put the responses in each batch in ascending order, then the median of each is the middle value as given below. (Obviously you could determine the middle value of three numbers without writing them down.)

Sample	Ordered responses			Median
A	1	2	4	2
B	1	4	5	4
C	1	4	4	4
D	1	3	5	3
E	1	1	1	1
F	3	5	5	5

Solution to Activity 8

The most noticeable, and most important, change is that, as the sample size increases, the proportion of samples with median 3 increases, whilst the proportions with medians 1, 2, 4 and 5 decrease.

For $n = 15$, already over half of the samples (actually about 0.574 of them) have median 3, and for $n = 45$ this proportion has risen even higher, to 0.827. For $n = 135$, nearly all the samples (a proportion of 0.981) have median 3.

Solution to Activity 9

The value with the largest proportion is the one with the longest vertical bar. This value is 4. (The proportion of the samples with median response 4 is 0.324.)

Solution to Activity 10

For each sample size n pictured, the proportion of the samples of size n with median 3 is as follows:

n	Proportion
3	0.296
5	0.365
9	0.467
15	0.574
25	0.692
45	0.827
75	0.921
135	0.981

Thus those sample sizes for which this proportion is larger than 60% (i.e. 0.6) are 25, 45, 75 and 135.

Solution to Activity 11

To choose strata, you need information that is both related to the subject under investigation and available for all individuals in the population before the survey starts. The only information that is mentioned as being available for all customers is name, address and quantity of seed bought. A customer's address is likely to be related to the geographical location where the customer grew the seeds, and satisfaction with the results might well be related to location because climate varies with location. Therefore, it would make sense to stratify in terms of geographical region. You might also have felt that a customer's satisfaction might be related to the amount of seed bought; if so, that could also be used for stratification.

You may have suggested other criteria for stratification, and these may well be sensible, but remember that a variable used for stratification needs to be known for all the customers before the sample is chosen.

Solution to Activity 12

The random numbers in row **6** start 71 58 45 Hence we start with the 7th person listed in Table 11, Mary Abbott. She is the first person in the sample and we then include every eighth person until we reach the end of the list. From the table, the people in the sample are: Mary Abbott (Y), Arthur Mason (N), Jacqueline Groves (Y), Sean Daly (Y), Mary Dunn (Y), Raymond McCann (Y), Nicholas Kohler (N), Annette Clarke (N), Jaya Shah (Y), Darrell Paine (N), Dipak Shah (Y) and Muriel Uttley (Y). In this sample of 12, the number of bus users is eight, so the sample estimate of the percentage of bus users in the population is again $8/12 \simeq 66.7\%$.

Solution to Activity 13

The first two single digits in row **26** are 3 and 2, which correspond to Lytham Gardens and Kerrera Close.

Lytham Gardens has 12 electors, so we will select $12/3 = 4$ of these. Kerrera Close has 27 electors so we will select $27/3 = 9$ of these.

Starting in row **82**, the random number pairs are as follows. (The pairs corresponding to selected registration numbers are given in italics.)

82	77 69 59	62 <i>33</i> 99	26 67 95	72 77 16	02 <i>28</i> 96
83	08 19 98	26 68 06	02 05 57	21 73 55	<i>35</i> 07 79
84	50 83 92	60 44 28	52 83 25	<i>39</i> 83 60	92 71 <i>10</i>
85	<i>16</i> 89 30	82 48 70	63 82 71	48 72 82	77 37 56
86	<i>21</i> 41 74	65 <i>08</i> 73	82 94 72	<i>22</i> 67 92	34 74 33
87	99 08 47	77 43 94	<i>17</i> 07 76	57 93 68	61 <i>15</i> 97
88	<i>20</i> 02 69	...			

The selected registration numbers for Lytham Gardens (from registration numbers 1028–1039) are: (10)33 (10)28 (10)35 (10)39.

Those for Kerrera Close (from registration numbers 1001–1027) are: (10)10 (10)16 (10)21 (10)08 (10)22 (10)17 (10)07 (10)15 (10)20.

Thus the people in the survey and their bus usages are: Nicholas Kohler (N), John Kerr (N), Marie Clements (N), Samuel Bolton (N), Audrey Ho (N), Garry Dunn (Y), Matthew Goulding (Y), Mary Dixon (Y), Janet Goulding (Y), Mary Dunn (Y), Alan Smith (Y), Clare Christon (Y) and Eleanor Price (N).

In this sample of 13, the number of bus users is seven, so the sample estimate of the percentage of bus users in the population is $7/13 \simeq 53.8\%$.

Solutions to exercises

Solution to Exercise 1

There are many ways of using the table to choose such a sample. Perhaps the most straightforward method uses groups of three digits, working along the rows from a randomly chosen starting point much as you did for the other two target populations in Subsection 1.2.

For example, if the starting point is the beginning of row **49**, then this method will select the following labels:

960 026 826 022 069 990 967.

With this starting point, the individual 026 was repeated and had to be ignored the second time. There may have been a problem with repeated individuals in your sample, but this is quite unlikely with a small sample from a large population.

Solution to Exercise 2

(a) The nine labels selected are

26 25 37 27 02 15 51 87 18.

To obtain this sample it is necessary to use 11 digit pairs from the table, because the labels 26 and 27 are repeated.

(b) The 17 labels selected are

32 57 79 72 02 27 96 10 62 63 07 30 01 40 97 69 23.

This time there is no problem with repetition: 17 digit pairs are enough.

Solution to Exercise 3

(a) The first eight pairs of digits from row **5** in the range 01 to 86 are

50 76 11 36 13 84 32 72.

The following two tables show the people in this sample and analyse the sample by gender and occupation.

Name	Label	Gender	Occupation
Kapoor, Sashi	50	M	P
Thompson, Jack	76	M	P
Bramley, Max	11	M	P
Grant, Lynne	36	F	P
Cameron, Lynne	13	F	P
Woodhouse, Paul	84	M	M
Franks, Abraham	32	M	P
Shah, Anjali	72	F	S

	Male	Female	Total
Professional	4	2	6
Administrative	0	0	0
Secretarial	0	1	1
Manual	1	0	1
Total	5	3	8

(b) The sample and its analysis are shown in the following tables.

Name	Label	Gender	Occupation
Singh, Meera	73	F	S
Bidford, David	09	M	P
Archer, Simon	03	M	M
London, Fred	54	M	P
Crofts, Dennis	20	M	P
Andrews, Jean	02	F	P
Lupton, David	55	M	P
Jolly, Susan	49	F	S
James, Patricia	48	F	A
Hutton, Joan	46	F	S
Thompson, Anna	75	F	S
Harrison, Sheila	42	F	P

	Male	Female	Total
Professional	4	2	6
Administrative	0	1	1
Secretarial	0	4	4
Manual	1	0	1
Total	5	7	12

(c) We must select every ninth label starting at label 05. Hence the sample is as follows.

Name	Label	Gender	Occupation
Baker, Fred	05	M	P
Carter, Jane	14	F	P
Daley, Stuart	23	M	P
Franks, Abraham	32	M	P
Hare, Dorothy	41	F	P
Kapoor, Sashi	50	M	P
Menton, Christine	59	F	S
Ricardo, Dan	68	M	P
Trumpington, Pat	77	F	S
Yeo, Tara	86	F	A

The following is an analysis of the sample.

	Male	Female	Total
Professional	5	2	7
Administrative	0	1	1
Secretarial	0	2	2
Manual	0	0	0
Total	5	5	10

- (d) This time we must select every tenth label starting at label 08, giving the following sample.

Name	Label	Gender	Occupation
Best, John	08	M	P
Cluskie, Alex	18	M	P
Estover, Matthew	28	M	P
Greenson, Denise	38	F	A
James, Patricia	48	F	A
Masterton, Dick	58	M	P
Ricardo, Dan	68	M	P
Truscott, Karen	78	F	S

The following is an analysis of the sample.

	Male	Female	Total
Professional	5	0	5
Administrative	0	2	2
Secretarial	0	1	1
Manual	0	0	0
Total	5	3	8

Solution to Exercise 4

- (a) 0.099.
 (b) 0.168.
 (c) To have a median less than 3 the sample must have median 1 or 2. So the proportion of samples with median less than 3 is the sum of the proportions with medians 1 and 2. This is $0.099 + 0.168$, which equals 0.267.
 (d) Similar reasoning implies that this is the sum of the proportions of samples with medians 4 and 5. This is $0.266 + 0.001 = 0.267$.

Note that the following proportions sum to one, approximately. The digit 1 in the last decimal place is due to rounding in the calculations.

Proportion with median less than 3	0.267
Proportion with median 3	0.467
Proportion with median greater than 3	0.267
Sum	1.001

The sum would be expected to be equal to 1 because each sample median is either less than 3, equal to 3 or greater than 3.

Solution to Exercise 5

- (a) The population size is 1000, so the median is halfway between the 500th and 501st values. Counting in 500 from the ‘Much worse off’ end of the population, the 500th and 501st values both equal 2. Hence the median is 2.
- (b) In Figure A, the proportion of samples that give a median of 2 is very small. As the population in the table comes from a population with a median of 2, Figure A must be the sample that relates to a different population. Looking at Figures B and C, the median is far more predictable from Figure B than from Figure C, so Figure B must relate to the larger sample. Thus Figure B is for a sample of size 21 from the tabulated population, while Figure C is for the sample of size 7.

Solution to Exercise 6

- (a) Cluster sampling would be valid and appropriate, because the expenditure pattern of such households is unlikely to be related to geographical area.
- (b) It would be difficult to obtain a valid sampling frame as there is no simple way to identify which households own their home outright and which do not. Therefore quota sampling would have to be used.

Solution to Exercise 7

- (a) Two lists could be obtained from official records: all addresses, and all addresses with a registered motor vehicle. From this, a list of all addresses at which no vehicles are registered could be obtained.
- (b) Either a simple or a systematic random sample would be sufficient, particularly as this is just a pilot survey.

Solution to Exercise 8

- (a) The sizes of the three strata are Jersey Close: 22; Kerrera Close: 27; other three roads: 46, which together total $22 + 27 + 46 = 95$. A total sample of size 12 is required, so the numbers to take from each stratum are:

$$\text{Jersey: } \frac{22}{95} \times 12 \simeq 3, \quad \text{Kerrera: } \frac{27}{95} \times 12 \simeq 3, \quad \text{other: } \frac{46}{95} \times 12 \simeq 6.$$

These sample sizes add to 12.

- (b) Starting at the beginning of row **52**, the selected registration numbers for Jersey Close (977–1000) are: 986 993 982.

For Kerrera Close (1001–1027): 1008 1019 1021.

For the third stratum (1028–1074): 1047 1056 1067 1040 1031 1046.

Hence the electors in the sample and their bus usages are:

Margaret Donegan (N), Joan Mason (Y), Linda Jones (Y),
Mary Dixon (Y), Vince Edwards (Y), Matthew Goulding (Y),
Brian Ferguson (N), David McGregor (N), Mala Shah (N),
David Clarke (N), David Kerr (Y) and Angela Smith (Y).

- (c) Seven individuals in this sample of 12 people are bus users, so the sample estimate of the percentage of bus users in the population is

$$\frac{7}{12} \times 100\% \simeq 58.3\%.$$

In the target population of 95 electors, there are 49 people who used the bus in the previous week, so the true percentage of bus users is $49/95 \simeq 51.6\%$. Hence the sample estimate is reasonably close to the population value. (For a sample of 12, the only sample result that would be closer is when the sample contains six bus users, which is only one different from the number in the sample we selected.)

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover image: Minxlj/www.flickr.com/photos/minxlj/422472167/. This file is licensed under the Creative Commons Attribution-Non commercial-No Derivatives Licence <http://creativecommons.org/licenses/by-nc-nd/3.0/>

Introduction, cartoon (Tower of Pisa), www.causeweb.org

Figure 2 Taken from:

www.ons.gov.uk/ons/guide-method/census/2011/index.html and
<http://www.scotlandscensus.gov.uk/en/>

Figure 3 © 2012 Microsoft Corporation

Figure 4 Taken from: Google Images

Figure 6 © 2012 Microsoft Corporation

Figure 11 David Ayres

Figure 12 This file is licensed under the Creative Commons Attribution Licence <http://creativecommons.org/licenses/by/3.0/>

Subsection 1.2 figure, ‘A UK National Lottery machine’, taken from:
<http://www.mirror.co.uk/money/city-news/lottery-set-for-42billion-boost-as-operator-753620>

Subsection 3.1 cartoon (hard to quantify), LightBulb Cartoon

Subsection 4.1 photo of George Gallup: Gallup Inc.

Subsection 4.1 cartoon (margin of error): John Landers, www.causeweb.org

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.